7-7-2020

Two-Stage Conditional Density Estimation Based on Bernstein Polynomials

Guanjie Lyu
University of Windsor

Two-Stage Conditional Density Estimation Based on Bernstein Polynomials

by

Guanjie Lyu

A Thesis
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

Two-Stage Conditional Density Estimation Based on Bernstein Polynomials

by

Guanjie Lyu

APPROVED BY:

A. Ngom
School of Computer Science

A. Hussein
Department of Mathematics and Statistics

M. Belalia, Advisor
Department of Mathematics and Statistics

May 19, 2020

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this major paper has been published or submitted for publication.

I certify that, to the best of my knowledge, my major paper does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my major paper, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my major paper and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my major paper, including any final revisions, as approved by my major paper committee and the Graduate Studies office, and that this major paper has not been submitted for a higher degree to any other University or Institution.

Abstract

In this thesis, we propose a new nonparametric approach based on Bernstein polynomials to estimate the conditional density function. The proposed estimators have desired properties at the boundaries, and can outperform the kernel and local linear estimators in terms of Integrated Mean Square Error for an appropriate choice of the polynomials' order. The idea is constructing a two-stage conditional probability density function estimator based on Bernstein polynomials. Specifically, the Nadaraya-Watson (NW) and local linear (LL) conditional distribution function estimators were smoothed using Bernstein polynomials in the first stage. Secondly, the proposed estimators are obtained by differentiating the smoothed Bernstein NW and LL estimators.

Further, the asymptotic properties of these estimators are established such as asymptotic bias, variance and normality under mild regularity conditions.

Finally, a simulation study is carried out to assess the relative advantage of our estimators compared to other estimators. Also, the well known Old Faithful Geyser data were analyzed using the proposed estimators.

Acknowledgements

First of all, I would like to thank my supervisor Dr. Belalia for his constant support during the career of my master. He is patient and kind all the time. Under his supervision, I feel relaxed and energized, which make me finish this thesis smoothly. I am also very grateful to all the members in my defense committee. Particularly, I would like to thank Dr. Ngom from school of computer science for agreeing to be my external program reader. I am also very grateful to Dr. Hussein for reading my thesis and giving me professional suggestions and to Dr. Caron for chairing the defense.

Besides professors, I am grateful to my parents for their significant support. Without their support and care I could not finish my courses and thesis.

Finally, I would like to offer my sincere thanks to all students, faculty members and staff in the department of Mathematics and Statistics for the harmonic, friendly and positive studying and working environment. In particular, I was deeply impressed by Ms. Dina Labelle and Mrs. Rose Spences' kind help and positive working attitudes.

Contents

List of Figures

Chapter 1

Introduction and motivation

Conditional probability density functions indicate comprehensive information on the relationship between an outcome and some predictor random variables. So, conditional density functions play an important role in statistics. The estimation of conditional densities responds to two fundamental problems in statistics: finding the distribution underlying a data set and describing the relationships between the different variables. From this point of view, the conditional densities estimation is a richer problem than two problems which have been intensively studied:

- the estimation of densities, which is naturally included by the estimation of conditional densities by not considering any variable as an auxiliary, and
- the problem of regression, conditional density actually contains more information than the regression function, which is simply conditional expectation, since from conditional density, we can obtain the regression function, but the reverse is false.

Compared to the two above mentioned problem, the literature is much poorer to deal

with the problem of estimating conditional densities, while there is a high demand in many fields of application such as economics (Hall et al., 2004), medicine (Takeuchi et al., 2009), actuarial (Efromovich, 2010) among others.

Usually, if the conditional density function has a known form, then the estimation turns to estimate some parameters, this is so-called parametric method. However, for certain statistical problems, the selection of a parametric model adapted to the data processed is not always easy. For this reason, nonparametric estimation and inference methods are good alternatives for this type of data. In this thesis, we focus on using nonparametric methods to estimate the conditional density function. Several nonparametric approaches have been proposed to estimate conditional density, such as kernel density estimators (Rosenblatt, 1969; Hyndman et al., 1996) and different methodologies for the bandwidth selection (Fan and Yim, 2004; Hall et al., 2004); local linear estimators (Fan and Gijbels, 1996; Hyndman and Yao, 2002) and methods based on Bernstein polynomials (Vitale, 1975; Babu et al., 2002; Babu and Chaubey, 2006; Belalia et al., 2017; Belalia, 2016; Leblanc, 2009, 2010), among others.

Nonparametric methods were used initially to estimate univariate density function by introducing histogram or kernels method. Furthermore, the regression function was estimated nonparametrically by Nadaraya (1965) and Watson (1964). Using the same approach, but in the context of conditional density estimation, Rosenblatt (1969) proposed a nonparametric estimator through plug-in kernel methods, which became the famous Nadaraya-Watson estimator.

The Nadaraya-Watson conditional distribution function estimator suffers from an excessive bias in the boundaries region. Furthermore, the fact that this estimator is a step function (not continuous) makes its derivative impossible. Thus, this latter

does not came with an associated density, which is too strict to apply under many circumstances. To correct the bias, local linear estimator (cf. Fan and Gijbels, 1996, Section 2.3.1) was developed. However, the local linear estimator still lead to that step function.

To overcome the limitation above, recently, Based on Bernstein polynomials, Belalia et al. (2017) proposed a new two stage conditional distribution estimators, which smooth the Nadaraya-Watson and local linear estimators and outperform the existing local polynomial conditional distribution estimator (see, Hansen (2004) and Hall et al. (1999)) in term of integrated mean square error. The resulting estimators are continuous, differentiable, and have an associated density.

Nonparametric estimation methods based on Bernstein polynomials start with the work of Vitale (1975). In that work, a Bernstein estimator for probability density function was introduced. It is studied further by Babu et al. (2002) and many others. This approach seems preferable to the kernel method on the boundary properties, see Leblanc (2012b). Bernstein polynomials were then used by many other researchers. For example, Babu and Chaubey (2006) considered the multivariate distribution function, and Belalia (2016) discussed the properties of the multivariate distribution function. The work of Ghosal (2001) and Petrone (1999b,a) discussed the Bayesian approach based on Bernstein polynomial. More recently, Bernstein polynomials were used by Belalia et al. (2019) to provide a nonparametric estimator of the conditional density function with application to conditional distribution and regression functions estimation.

The focus of this thesis is to study the resulting conditional probability density function using Bernstein polynomials. Specifically, the smoothed version of the Nadaraya-

Watson and local linear estimators proposed in Belalia et al. (2017) are derived to obtain the conditional density estimators based on Bernstein polynomial. The reminder of this thesis is organized as follow: In Chapter 2, an overview of kernel nonparametric estimation method for statistical quantities such as, cumulative distribution and its associated density functions, the conditional mean and conditional density functions are summarized. Similarly, nonparametric estimation methods based on Bernstein polynomials will be discussed in Chapter 3. Finally, The main contribution of this thesis is presented in chapter 4. This include, presenting the two-stage Bernstein conditional density estimator, providing its asymptotic properties: such as asymptotic bias, variance and establishing the asymptotic normality. A simulation study is carried out to assess the performance of the proposed estimators compared to Nadaraya-Watson and local linear estimators. the proposed estimators were used to analyze the Old Faithful Geyser Data.

Chapter 2

Kernel estimation methods

2.1 Statistical model

Assume that we are observing n independent and identically distributed ($i.i.d.$) sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from a couple of random variable (X, Y). Let F be the joint cumulative distribution function (*cdf*) and f its associated density function. This joint density satisfies

$$\mathbb{P}\left[a \leq X \leq b, c \leq Y \leq d\right] = \int_a^b \int_c^d f(x, y)\, \mathrm{d}x\, \mathrm{d}y. \tag{2.1}$$

The marginal *cdf* of X and its associated density are denotes by G, and g respectively. The conditional density of Y given X can be calculated by the ratio of the joint density f to the marginal density g and is shown in the following formula

$$f_x(y) = \frac{f(x, y)}{g(x)}, \tag{2.2}$$

where the value of $g(x)$ is fixed and greater than 0. The probability that Y will fall between a and b given that $X = x$ is obtained by

$$\mathbb{P}\left[a \leq Y \leq b \mid X = x\right] = \int_a^b f_x(y)\,\mathrm{d}y.$$

Before moving to the conditional density function estimation, let us review some most commonly used nonparametric estimation method for cumulative distribution and its associated density functions. The next section deals with the simplest nonparametric method to estimate a density f of a random variable X.

2.2 Univariate kernel density estimation

2.2.1 Histogram

The oldest and most widely used nonparametric estimator of a density f from an independent and identically distributed (*i.i.d.*) sample $X_1, \ldots, X_n$ is the *histogram*. The idea consists in aggregating the observation in intervals of the form $[x_0, x_0 + h)$ and then use their relative frequency to approximate the density at $x \in [x_0, x_0 + h)\,, f(x)$ by the estimate of

$$\begin{aligned} f(x_0) &= F'(x_0) \\ &= \lim_{h \to 0^+} \frac{F(x_0 + h) - F(x_0)}{h} \\ &= \lim_{h \to 0^+} \frac{\mathbb{P}[x_0 < X < x_0 + h]}{h}. \end{aligned}$$

Precisely, given an origin x_0 and a bin width $h > 0$, the histogram builds a piecewise

constant function in the intervals $\{B_\ell = [x_0 + \ell h, x_0 + (\ell + 1)h), \ell \in \mathbb{Z}\}$ by counting the number of sample points inside each of them. These constant-length intervals are also denoted *bins*. The fact that they are of constant length h is important, since it allows to standardize by h in order to have relative frequencies per length in the bins. For a given point $x \in B_\ell$, the histogram is defined as

$$\hat{f}_{nh}(x) = \frac{1}{nh}(\text{numbers of } X_i \text{ in same bin as } x) = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{I}_{\{X_i \in B_\ell, x \in B_\ell\}},$$

where n is the number of observations.

The intuition of this density estimator is that the histogram assign equal density value to every point within the bin. Note that, to construct the histogram, we have to choose both an original x_0 and a bin width h. The choice of h, primarily, controls the amount of smoothing inherent in the procedure.

The histogram may be affected by three effects, the choice of origin, the coordinates and the smooth parameter, thus though the histogram is a good estimate for large sample size, it is difficult to get a high precision estimate for small sample size as illustrated in Figure 2.1.

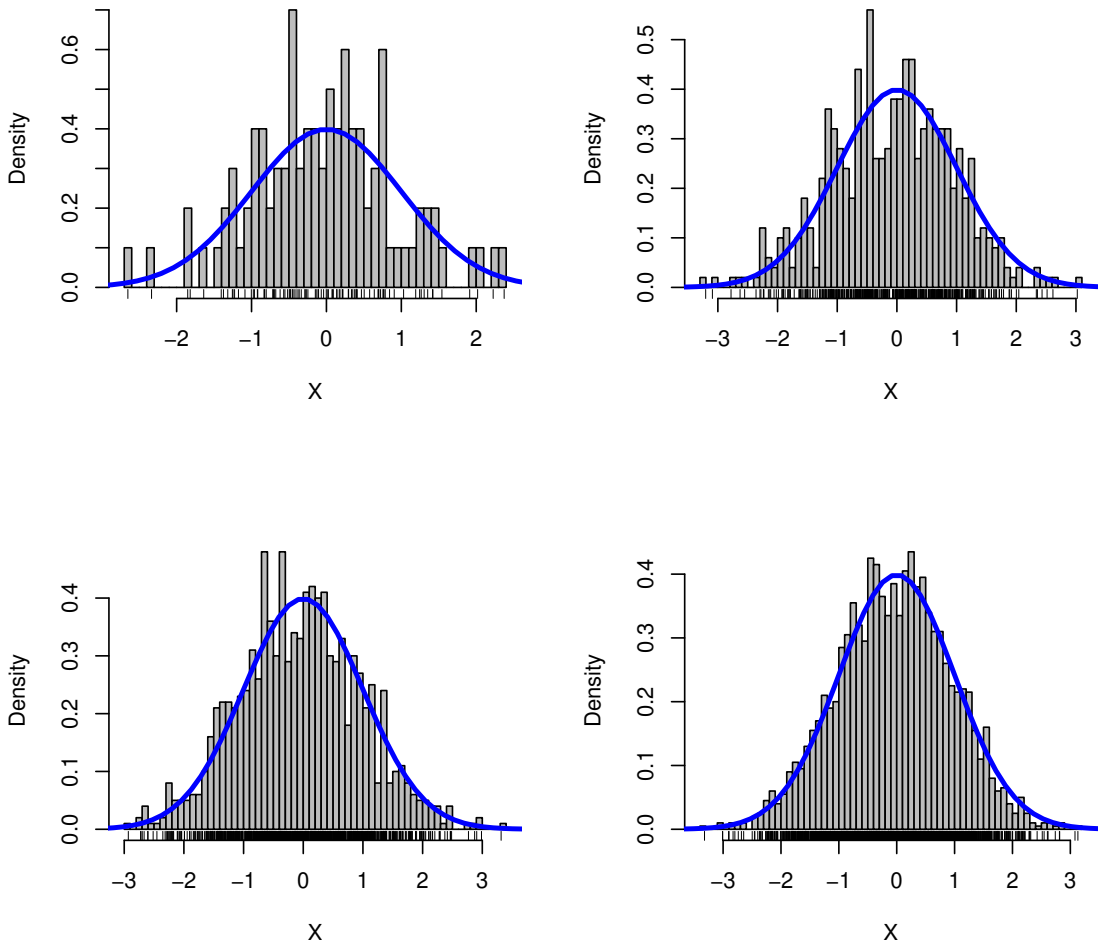


Figure 2.1: The histogram density estimator for the standard normal density. The sample size is $n = 100, 500, 1000, 2000$.

2.2.2 Kernel estimation methods construction

In this section, we review some of most widely used nonparametric estimation kernel based methods. The approach was introduced to estimate statistical quantities such as, *cdf* and its associated density, regression and quantile functions, among others. For

more details about nonparametric techniques, the reader is referred to the excellent monograph by Li and Racine (2007). In what follows, we will focus on density function estimation.

The kernel density function estimator was proposed by Rosenblatt (1956) based on the idea of that deriving the empirical cumulative function. Let $F_n : \mathbb{R} \longrightarrow [0, 1]$ the empirical cumulative distribution function, which a nonparametric way to estimate the *cdf* G, given by

$$F(x) = \mathbb{P}\left(X \leq x\right). \tag{2.3}$$

Starting from a *i.i.d* sample $(X_1, X_2, \ldots, X_n)$ drawn from F, intuitively, the empirical cumulative distribution function at point x is the number of observation $X_i, i = 1, 2, \ldots n$ fall before x, formally, F_n is defined as

$$F_n(x) = \frac{1}{n}\{\text{numbers of } X_i's \leq x\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x) \tag{2.4}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Some of the most important asymptotic properties of F_n, such as asymptotic bias, variance, and distribution limit are postponed in A.2.

To avoid the dependence of the histogram estimator on the origin x_0, *the moving histogram* or *naive density estimator* was introduced as alternative. Starting from the definition of a probability density function (*pdf*) denoted as f, we have

$$f(x) = \frac{d}{\mathrm{d}x}F(x), \tag{2.5}$$

then an estimate of $f(x)$ can be obtained as

$$\hat{f}_{nh}(x) = \frac{d}{dx}F_n(x) = \lim_{h\to 0^+} \frac{F_n(x+h) - F_n(x-h)}{2h}, \tag{2.6}$$

where $h > 0$ is a small positive increment. By substituting (2.4) into (2.6), we have

$$\begin{aligned}\hat{f}_{nh}(x) &= \frac{1}{2nh}\{\text{numbers of } X_i's \text{ falling into } [x-h, x+h]\} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} \mathbb{I}_{\{x-h<X_i<x+h\}}. \end{aligned} \tag{2.7}$$

This estimator is also called naive density estimator, and is illustrated in Figure 2.2 with the effect of the bandwidth parameter h.

The properties of $\hat{f}_{nh}(x)$ as a random variable follows by observing that

$$\sum_{i=1}^{n} \mathbb{I}_{\{x-h<X_i<x+h\}} \sim \text{Binomial}(n, p_{x,h}),$$

where

$$p_{x,h} := \mathbb{P}[x-h < X < x+h] = F(x+h) - F(x-h).$$

Therefore, employing the bias and variance expressions of a binomial, it follows:

Theorem 2.1. *The expectation and variance of $\hat{f}_{nh}(x)$ are given, respectively by*

- *The expectation*

$$\mathbb{E}[\hat{f}_{nh}(x)] = \frac{F(x+h) - F(x-h)}{2h},$$

- *The variance*

$$\mathbb{V}\mathrm{ar}[\hat{f}_{nh}(x)] = \frac{F(x+h) - F(x-h)}{4nh^2} - \frac{\left(F(x+h) - F(x-h)\right)^2}{4nh^2}.$$

Proof of Theorem 2.1: A detailed proof of this theorem is given in A.3.

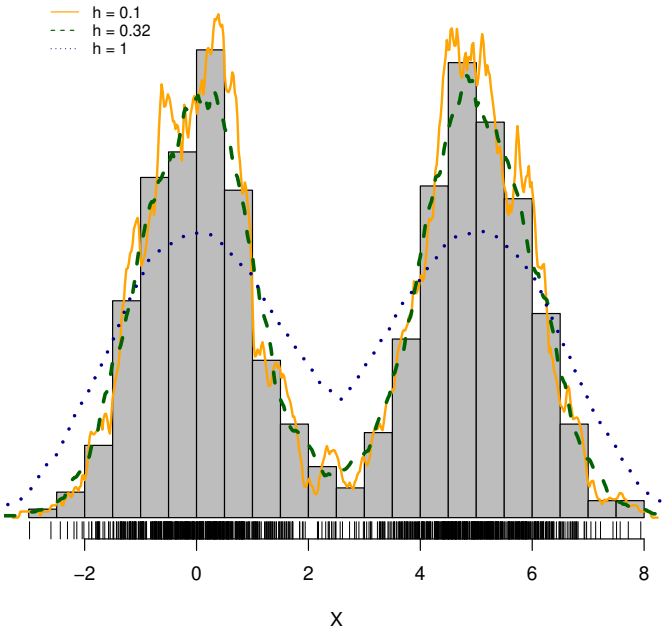


Figure 2.2: Illustration of naive density estimator with three value of the bandwidth parameter $h = 1$ (orange line), $h = 0.32$ (green dashed line), and $h = 1$ (blue dotted line).

We follow this idea to extend the naive density estimator to the general weight function estimator Silverman (1986). The general weight function estimator is the convolution of the empirical distribution function and a weight function. But we firstly need the introduction of Dirac Delta function as a useful tool for the derivation of

this class of estimators.

The Dirac delta function is the derivative of the Heaviside function defined as

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ \\ 1 & \text{if } x > 0 \end{cases},$$

we denote it as $\delta(x)$. The Dirac delta function is actually a distribution which satisfies following properties

i.
$$\delta(x) = \begin{cases} 0 & \text{if } x \neq 0 \\ \\ \infty & \text{if } x = 0 \end{cases},$$

ii.
$$\int_{-\infty}^{\infty} \delta(x)\,\mathrm{d}x = 1,$$

iii.
$$x\delta(x) \equiv 0,$$

since the $\delta(x)$ is zero for $x \neq 0$ and suppose $f(x)$ is a continuous function, then properties of delta function allow us to write

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\delta(x)\,\mathrm{d}x &= \int_{-\infty}^{\infty} f(0)\delta(x)\,\mathrm{d}x \\ &= f(0), \end{aligned}$$

thus the derivative of empirical distribution function can be rewritten as

$$\hat{f}_{nh}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - X_i).$$

The general weight function estimator denoted as $\hat{f}_w(x)$ is given by

$$\begin{aligned}\hat{f}_w(x) &= \int_{-\infty}^{\infty}\hat{f}_n(x-t)w(t)\,\mathrm{d}t \\ &= \int_{-\infty}^{\infty}\frac{1}{n}\sum_{i=1}^{n}\delta(x-t-X_i)w(t)\,\mathrm{d}t \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\delta(x-t-X_i)w(t)\,\mathrm{d}t \\ &= \frac{1}{n}\sum_{i=1}^{n}w(x-X_i), \end{aligned} \tag{2.8}$$

where $w(\cdot)$ is a continuous function satisfying the following conditions

$$\int_{-\infty}^{\infty} w(x,t)\,\mathrm{d}t = 1,$$

and $w(x,t) \geq 0$ for all x and t.

However, when

$$w(x,t) = \frac{1}{h}K\left(\frac{t-x}{h}\right),$$

where $K(\cdot)$ satisfies the following regularity conditions:

$$\int_{-\infty}^{\infty} K(v)dv = 1, \quad K(v) = K(-v), \text{ and } \int_{-\infty}^{\infty} v^2K(v)dv = \kappa_2 > 0, \tag{2.9}$$

the resulting weight function estimator is called kernel density estimator, and it is

given by

$$\hat{f}_{nh}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{2.10}$$

Note that if the kernel is reduced to the rectangular function defined as

$$K(x) = \begin{cases} 1/2 & \text{if } |x| < 1 \\ 0 & \text{otherwise} \end{cases},$$

the kernel density estimator is reduced to the naive density estimator. Some of the asymptotic properties of $\hat{f}_{nh}(x)$ are given in the following theorem.

Theorem 2.2. *Let $X_1, ..., X_n$ denote i.i.d. observations having a three-times differentiable pdf $f(x)$, and let $f^{(s)}(x)$ denote the s^{th} order derivative of $f(x)(s = 1, 2, 3)$. Let x be an interior point in the support of X. Assume that the kernel function $K(\cdot)$ is bounded and satisfies* (2.9)*. Also, as $n \to \infty$, $h \to 0$ and $nh \to \infty$, then, the estimator* (2.10) *satisfies;*

$$\begin{aligned} \text{MSE}(\hat{f}_{nh}(x)) &= \frac{h^4}{4}\left[\kappa_2 f^{(2)}(x)\right]^2 + \frac{\kappa f(x)}{nh} + o(h^4 + (nh)^{-1}) \\ &= O(h^4 + (nh^{-1})), \end{aligned} \tag{2.11}$$

where $\kappa_2 = \int v^2 K(v)dv$ and $\kappa = \int K^2(v)dv$.

The proof of Theorem 2.2 is given in Appendix A.4.

An intuitive construction of the kernel density estimator defined by Equation (2.10) of the sample $X = 65, 75, 67, 79, 81, 91$, is depicted in Figure 2.3. This construction can be done as follows: we place a normal kernel with standard deviation 5.5 (indicated by the red dashed lines) on each of the data points $x_i, i = 1, 2, \ldots, 6$. The kernels are

summed to make the kernel density estimate (solid blue curve). The smoothness of the kernel density estimate is evident compared to the discreteness of the histogram, as kernel density estimates converge faster to the true underlying density for continuous random variables. Also, an application on the Faithful Geyser dataset to estimate the waiting time before the next eruption is illustrated in in Figure 2.4. Finally, the effect of the bandwidth parameter is shown in using a sample of size $n = 100$ drawn from the standard normal distribution.

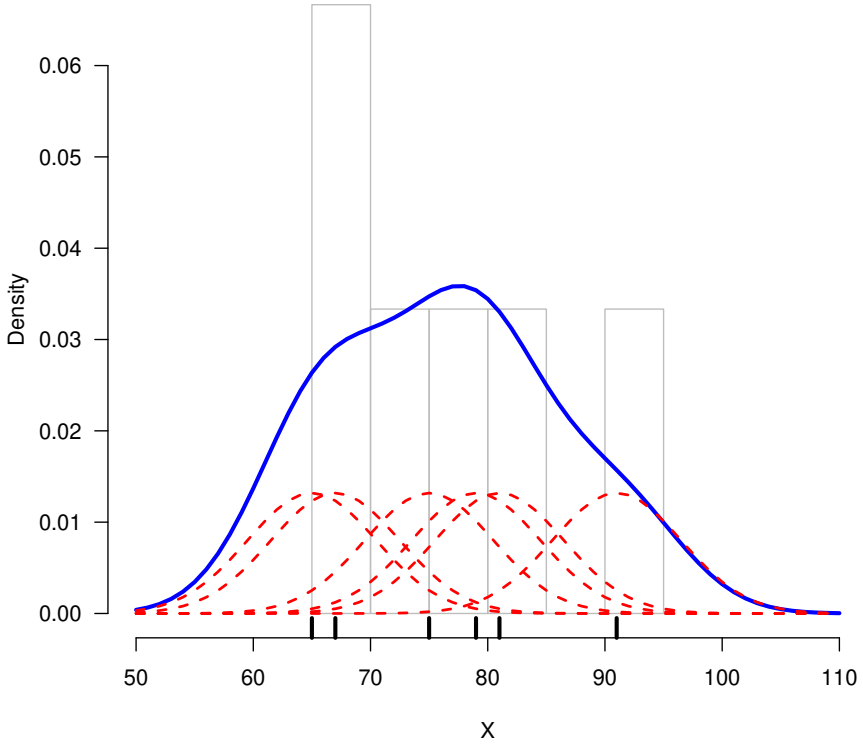


Figure 2.3: Kernel density estimate constructed using the same data. The six individual kernels are the red dashed curves, the kernel density estimate the blue curves. The data points are the rug plot on the horizontal axis.

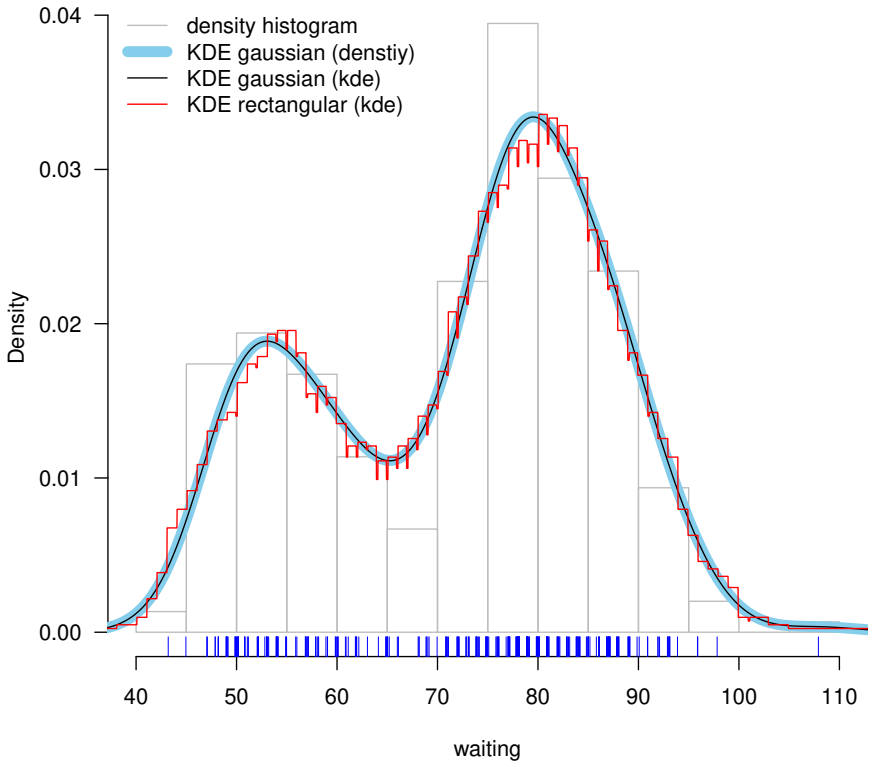


Figure 2.4: Kernel density estimation (KDE) of the waiting time before the next eruption.

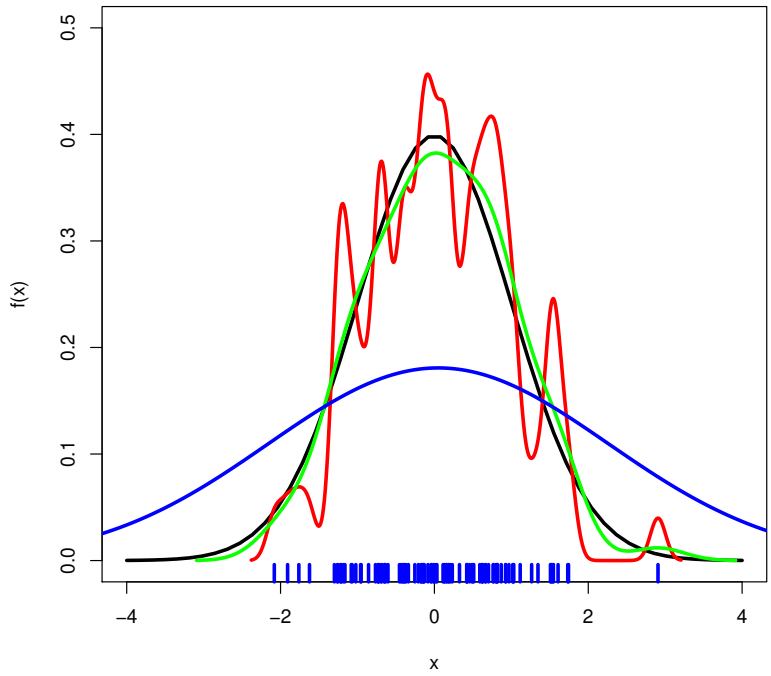


Figure 2.5: Kernel density estimate (KDE) with different bandwidths of a random sample of 100 points from a standard normal distribution. Black: true density (standard normal). Red: KDE with $h = 0.1$. Green: KDE with $h = 0.337$. Blue: KDE with $h = 2$.

2.3 Multivariate kernel density estimation

Kernel density estimation discussed above can be generalized to estimate multivariate densities $f \in \mathbb{R}^d$ in a straightforward way. Suppose now we have observations $(\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n})$, where each of the observations is a d-dimensional vector $\mathbf{X_i} = (X_{i1}, X_{i2}, \ldots, X_{id})^T$. The multivariate kernel density estimator at point $\mathbf{x} =$

$(x_1, x_2, ..., x_d)^T$ is defined as

$$\hat{f}_{n\mathbf{H}}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|^{1/2}} \sum_{i=1}^{n} \mathcal{K}\left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{X}_i)\right), \tag{2.12}$$

with $\mathcal{K}$ denoting a multivariate kernel function, a $d-$variate density that is (typically) symmetric and unimodal at $\mathbf{0}$, and that depends on the bandwidth matrix $\mathbf{H}$, a $d \times d$ symmetric and positive definite matrix.

A common simplification is to consider a diagonal bandwidth $\mathbf{H} = \text{diag}(h_1^2, \ldots, h_d^2)$, which leads to the multivariate kernel density estimator employing product kernels:

$$\hat{f}_{n\mathbf{h}}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1 ... h_d} K\left(\frac{X_{i1} - x_1}{h_1}\right) K\left(\frac{X_{i2} - x_2}{h_2}\right) \ldots K\left(\frac{X_{id} - x_d}{h_d}\right), \tag{2.13}$$

where $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})^\top$ and $\mathbf{h} = (h_1, \ldots, h_d)^\top$ is the vector of bandwidths.

To illustrate the usefulness of the bivariate kernel density estimator, the joint density of the duration and waiting time in the Faithful Geyser dataset is plotted in Figure 2.6.

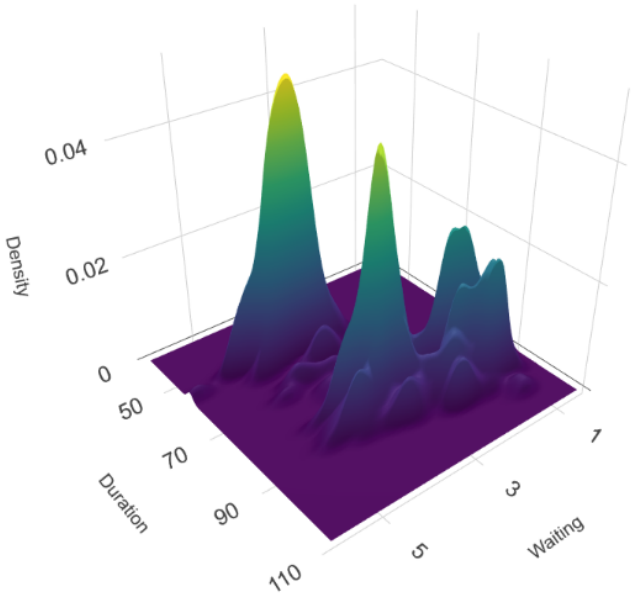


Figure 2.6: Bivariate kernel density estimation of duration and waiting time of faithful geyser data.

2.4 Kernel conditional density estimation

In this section the main kernel nonparametric method for conditional density estimates is presented. Indeed, the Nadaraya-Watson estimator is presented in subsection 2.4.1, and the Local Linear estimator in the section 2.4.2.

2.4.1 Nadaraya-Watson Estimator

To help motivate the construction of the Nadaraya-Watson conditional density estimator, we first discuss the regression function $m : \mathbb{R} \to \mathbb{R}$ estimator. Due to its

definition, $m(\cdot)$ can be rewritten as

$$m(x) = \mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y \frac{f(x,y)}{g(x)} \, \mathrm{d}y. \tag{2.14}$$

This expression shows an interesting point: the regression function can be computed from the joint density f and the marginal g. Therefore, given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, a nonparametric estimate of m can be obtained by replacing the previous densities by their kernel density estimators. We can therefore define the estimator of m as

$$\begin{aligned}
\frac{\int y \, \hat{f}_{n\mathbf{h}}(x,y) \, \mathrm{d}y}{\hat{g}_{nh_x}(x)} &= \frac{\int y \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - X_i) K_{h_y}(y - Y_i) \, \mathrm{d}y}{\frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - X_i)} \\
&= \frac{\frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - X_i) \int y K_{h_2}(y - Y_i) \, \mathrm{d}y}{\frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - X_i)} \\
&= \frac{\frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - X_i) Y_i}{\frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - X_i)} \\
&= \sum_{i=1}^{n} \frac{K_{h_x}(x - X_i)}{\sum_{i=1}^{n} K_{h_x}(x - X_i)} Y_i.
\end{aligned}$$

The resulting estimator the so-called Nadaraya–Watson estimate of the regression function:

$$\hat{m}_n(x) := \sum_{i=1}^{n} \frac{K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)} Y_i = \sum_{i=1}^{n} w_i^{\mathrm{NW}}(x) Y_i, \tag{2.15}$$

where $h_x = h$, and

$$w_i^{\mathrm{NW}}(x) := \frac{K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}.$$

For a visual aspect of $\hat{m}_n$, the Example 1 of Hall et al. (1999) and also considered by Veraverbeke et al. (2014) is used for illustration. Specifically, consider the case where

$$Z_i = 2\sin(\pi X_i) + \epsilon_i, \qquad i = 1, ..., n, \tag{2.16}$$

and where $\{X_i\}$ and $\{\epsilon_i\}$ are two independent sequences of independent random variables each having density $1 - |x|$ on $[-1, 1]$. Figure 2.7 displays a typical data set generated from model (2.16) using $n = 200$ observations with the associated regression curve $y = 2\sin(\pi x)$.

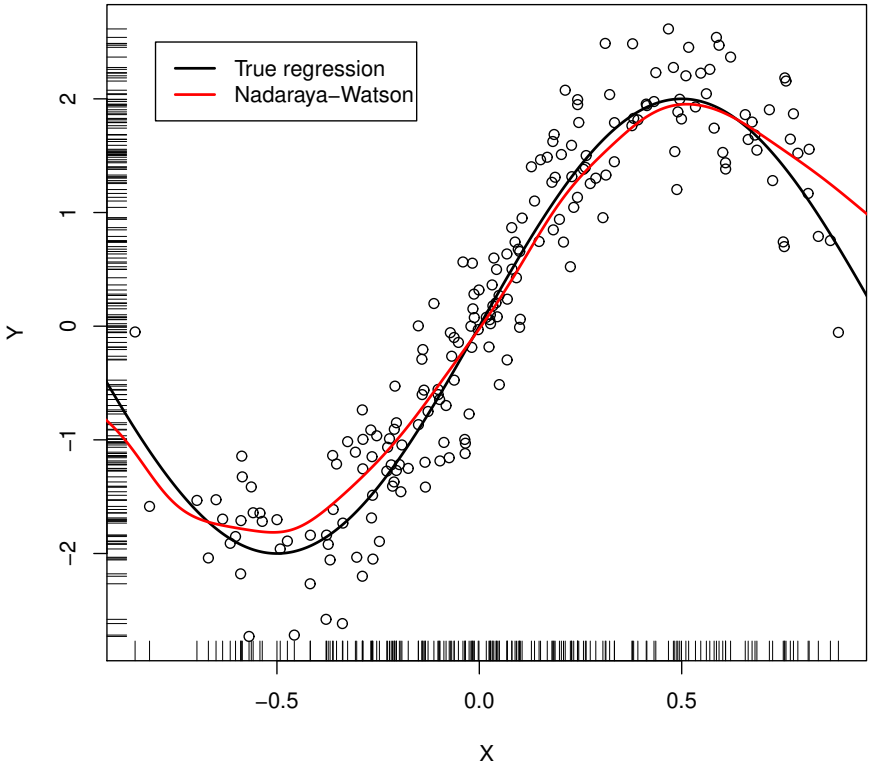


Figure 2.7: Typical data set generated from model (2.16) using $n = 200$ and true mean curve $y = 2\sin(\pi x)$. The bandwidth parameter is $h = 0.1$.

Now we can follow the idea in Stone (1977) to construct the Nadaraya-Watson conditional distribution estimator. In fact, the conditional cumulative distribution can be rewritten as the conditional mean of $\mathbb{I}(Y \leq y)$ given $X = x$, namely,

$$F_x(y) = \mathbb{P}\left[Y \leq y | X\right] = \mathbb{E}\left[\mathbb{I}(Y \leq y) | X\right].$$

This naturally suggests to use a regression approach to estimate F_x and is the basis for most of the work done so far on conditional CDF nonparametric estimation. For

instance, using the same approach as Nadaraya (1964, 1965) and Watson (1964), we can estimate $F_x(y)$ by

$$\hat{F}_{x,h}(y) = \frac{\sum_{i=1}^{n} K_{h_x}(x - X_i)\, \mathbb{I}(Y_i \leq y)}{\sum_{j=1}^{n} K_{h_x}(x - X_j)},$$

which can also be written as

$$\hat{F}_{x,h}(y) = \sum_{i=1}^{n} w_i^{\mathrm{NW}}(x)\, \mathbb{I}(Y_i \leq y), \tag{2.17}$$

where $K_h(x) = h^{-1}K(x/h)$, K is a kernel function, $h = h_x$ is the smoothing bandwidth and the definition of the weights w_i is obvious.

Similarly, by using the definition of the conditional density function given by Equation (2.2), the Nadaraya-Watson estimator of the conditional density f_x, can be obtained as

$$\hat{f}_{x,\mathbf{h}}(y) = \frac{\hat{f}_{n\mathbf{h}}(x,y)}{\hat{g}_{nh_x}(x)} = \frac{\sum_{i=1}^{n} K_{h_x}(X_i - x)\, K_{h_y}(Y_i - y)}{\sum_{i=1}^{n} K_{h_x}(X_i - x)}. \tag{2.18}$$

Theorem 2.3. *Assuming the the conditional density function $f_x(y)$ has bounded and continuous second order derivative with respect to y, we have*

$$\mathbb{B}\mathrm{ias}\left(\hat{f}_{x,\mathbf{h}}(y)\right) = \frac{h_y^2 \kappa_2}{2} \frac{\partial^2 f_x(y)}{\partial y^2} + o(h_y^2),$$

where $\kappa_2 = \int t^2 K_y(v)dv$. Also, we have

$$\mathbb{V}\mathrm{ar}\left(\hat{f}_{x,\mathbf{h}}(y)\right) = \frac{\kappa^2}{nh_x h_y} \frac{f_x(y)}{g(x)} + o((nh_x h_y)^{-1}),$$

where $\kappa = \int K_y(v)dv$, $g(\cdot)$ *is the univariate pdf. And*

$$\text{AMSE}\left(\hat{f}_{x,\mathbf{h}}(y)\right) = \frac{h_y^4\kappa_2^2}{4}\left(\frac{\partial^2 f_x(y)}{\partial y^2}\right)^2 + \frac{\kappa^2}{nh_xh_y}\frac{f_x(y)}{g(x)},$$

provided that the bandwidth h_x *and* h_y *converge to zero in such a way that* $nh_xh_y \to \infty$.

2.4.2 Local Linear Estimator

In this section, we will discuss the limitation of NW regression estimator and introduce an improved estimator called local linear estimator.

Consider a simple case of regression function such as $Y_i = \alpha + X_i\beta$, the performance of this regression function will depend on the marginal distribution of the X_i. If they are not spaced at uniform distances, then $\hat{m}_n(x) \neq m(x)$. One way to see the source of the problem is to consider the nonparametric equation $\mathbb{E}(X_i - x|X_i = x) = 0$. The numerator of the NW estimator is

$$\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(X_i - x),$$

but this is non-zero. Another problem of NW estimator occurs at the boundary of the support. In fact, the estimator is inconsistent at the boundary. To solve these problems, the local polynomial estimator is introduced, see Fan and Gijbels (1996). The motivation for the local polynomial fit comes from attempting to find an estimator $\hat{m}_n$ of m that minimizes the residual sum of squares (RSS)

$$\sum_{i=1}^{n}\left(Y_i - \hat{m}(X_i)\right)^2 \tag{2.19}$$

without assuming any particular form for the true m. We use Taylor expansion

$$m\left(X_i\right) \approx m(x) + m'(x)\left(X_i - x\right) + \ldots + \frac{m^{(p)}(x)}{p!}\left(X_i - x\right)^p \tag{2.20}$$

to induce a local parametrization on m with p^{th} order.

Then we replace (2.20) into (2.19), we have

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{p}\beta_j\left(X_i - x\right)^j\right)^2,$$

where $\beta_j = \frac{m^{(j)}(x)}{j!}$. In this way, we eliminate the $m(\cdot)$, and turn to estimate $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)$. The final touch is to use a weighted least squares by the kernel function to estimate the $\boldsymbol{\beta}$, which is

$$\hat{\boldsymbol{\beta}}_h = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{p}\beta_j\left(X_i - x\right)^j\right)^2 K_{h_x}\left(X_i - x\right). \tag{2.21}$$

We denote

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{bmatrix},$$

and

$$\mathbf{W} = \operatorname{diag}(K_{h_x}(X_1 - x), \ldots, K_{h_x}(X_n - x)), \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}.$$

Then we can re-express (2.21) as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_h &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})' \mathbf{W} (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\ &\Rightarrow 2(-\mathbf{X}')\mathbf{W}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \\ &= \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}.\end{aligned}$$

The estimate for $m(x)$ can be rewritten as

$$\begin{aligned}\hat{m}_n(x) &= \mathbf{e_1}'\hat{\boldsymbol{\beta}}_h \\ &= \mathbf{e_1}'\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \\ &= \sum_{i=1}^{n} W_i^p(x) Y_i,\end{aligned}$$

where

$$W_i^p(x) = \mathbf{e}_1'\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{e}_i,$$

and $\mathbf{e}_i$ is the i_{th} standard basis vector. We can notice that the local polynomial estimator is a weighted linear combination with the responses, just the same as the Nadaraya-Watson estimator. In fact, when $p = 0$, the local polynomial estimator is the NW estimator, also called *local constant estimator*. When $p = 1$, we have

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix},$$

and

$$\mathbf{W} = \text{diag}(K_{h_x}(X_1 - x), \ldots, K_{h_x}(X_n - x)), \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}.$$

Let

$$S_j = \sum_{i=1}^{n} K_{h_x}(x - x_i)(x - x_i)^j, \qquad \text{for} \quad j = 0, 1, 2.$$

And

$$V_0 = \sum_{i=1}^{n} K_{h_x}(x - x_i)Y_i, \quad V_1 = \sum_{i=1}^{n} K_{h_x}(x - x_i)(x - x_i)Y_i.$$

Then we have

$$\begin{aligned} \hat{m}_n(x) &= \mathbf{e}_1' \left(\begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix}' \begin{bmatrix} K_{h_x}(x - x_i) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K_{h_x}(X_n - x) \end{bmatrix} \begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix} \right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{Y} \\ &= \mathbf{e}_1' \begin{bmatrix} S_0 & S_1 \\ S_1 & S_2 \end{bmatrix}^{-1} \begin{bmatrix} V_0 \\ V_1 \end{bmatrix}. \end{aligned}$$

According to the inverse formula of partitioned matrix

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix},$$

thus, we have

$$\begin{aligned}\hat{m}_n(x) &= \left(S_0 - S_1 S_2^{-1} S_1\right)^{-1} \left(V_0 - S_1 S_2^{-1} V_1\right) \\ &= \frac{\sum_{i=1}^{n} w_i^{\mathrm{LL}}(x) Y_i}{\sum_{i=1}^{n} w_i^{\mathrm{LL}}(x)},\end{aligned} \tag{2.22}$$

where $w_i^{\mathrm{LL}}(x) = K_{h_x}(x - x_i)[1 - S_1 S_2^{-1}(x - x_i)]$.

Note that, one can handle the local linear regression estimator to get a conditional distribution function estimator in the same way as the NW regression estimator, which lead to

$$\hat{F}_{x,h}(y) = \frac{\sum_{i=1}^{n} w_i^{\mathrm{LL}}(x) \mathbb{I}\left(Y_i \leq y\right)}{\sum_{i=1}^{n} w_i^{\mathrm{LL}}(x)},$$

where $w_i^{\mathrm{LL}}(x)$ is the same as in (2.22).

As described in Fan and Gijbels (1996), and following the same strategy as for the NW conditional density estimator, the local linear (LL) conditional density function estimator can be stated as,

$$\hat{f}_{x,\mathbf{h}}(y) \approx \mathbb{E}\left(K_{h_y}(Y - y) | X = x\right) = \frac{\sum_{i=1}^{n} w_i^{\mathrm{LL}}(x) K_{h_y}(Y_i - y)}{\sum_{i=1}^{n} w_i^{\mathrm{LL}}(x)}, \tag{2.23}$$

where $K_h(\cdot)$ is a kernel function as previously.

Theorem 2.4. *Under Condition* 2 *in Fan and Gijbels (1996, Section* 6.6*), we have*

$$\mathbb{B}\mathrm{ias}\left(\hat{f}_{x,\mathbf{h}}(y)\right) = \frac{h_x^2 \mu_2}{2} \frac{\partial^2 f_x(y)}{\partial x^2} + \frac{h_y^2 \mu_K}{2} \frac{\partial^2 f_x(y)}{\partial y^2} + o(h_x^2 + h_y^2),$$

where $\mu_K = \int t^2 K_y(v)dv, \mu_j = \int t^j K_x(v)dv$. *Also, we have*

$$\mathbb{V}\text{ar}\left(\hat{f}_{x,\mathbf{h}}(y)\right) = \frac{\nu_K \nu_0}{nh_x h_y}\frac{f_x(y)}{g(x)} + o((nh_x h_y)^{-1}),$$

where $\nu_K = \int \{K_y(v)\}^2 dv, \nu_j = \int t^j \{K_x(v)\}^2 dv$, $g(x)$ *is the univariate PDF. And*

$$\begin{aligned}\text{AMSE}\left(\hat{f}_{x,\mathbf{h}}(y)\right) = \frac{\nu_K \nu_0}{nh_x h_y}\frac{f_x(y)}{g(x)} + \frac{h_x^4 \mu_2^2}{4}\left(\frac{\partial^2 f_x(y)}{\partial x^2}\right)^2 + \frac{h_y^4 \mu_K^2}{4}\left(\frac{\partial^2 f_x(y)}{\partial y^2}\right)^2 \\ + \frac{h_x^2 h_y^2 \mu_2 \mu_K}{2}\frac{\partial^2 f_x(y)}{\partial x^2}\frac{\partial^2 f_x(y)}{\partial x^2},\end{aligned}$$

provided that the bandwidth h_x *and* h_y *converge to zero in such a way that* $nh_x h_y \to \infty$.

Chapter 3

Bernstein estimation methods

3.1 Bernstein estimation methods

Nonparametric estimation methods based on Bernstein polynomials (Lorentz, 1986, cf.) are known by their optimal properties in terms of the mean square error (MSE). In addition, these estimation procedures behave in an interesting manner in the boundaries of the support of the distribution function or of its density, in particular the absence of bias at the border points.

The story started in 1913 when Sergeï Bernstein sought to give a constructive and probabilistic demonstration of Weierstrass' classical theorem, on the approximation of continuous functions over closed and bounded intervals, which can be stated as follow.

Theorem 3.1. *(Weierstrass Theorem). Let* $f : [a, b] \to \mathbb{R}$ *be a continuous real-*

function. Given $\epsilon > 0$, there exists a polynomials $Q_n(x)$ satisfying

$$\textit{for all} \quad x \in [a, b], \quad |f(x) - Q_n(x)| < \epsilon.$$

It is in this perspective that Sergeï Bernstein introduced a family of polynomials, which will bear his name later, an example of these polynomials are depicted in Figure 3.1a, and its definition is as follows.

Definition 3.1 (Bernstein polynomials)**.** *For $m \in \mathbb{N}$ and $0 \leq k \leq m$, the Bernstein polynomials $P_{m,k}$ of degree m are defined as*

$$P_{m,k}(x) = \binom{m}{k} x^k (1-x)^{m-k}, \quad k = 0, 1, 2, \ldots, m.$$

for $x \in [0, 1]$.

These polynomials have analytical-probabilistic properties, which until today attract many probabilistic and statisticians combined. We cite some of them by way of illustration.

Proposition 1. *(Properties) Bernstein polynomials have the following properties:*

(i) Partition of unity:

$$\sum_{k=0}^{m} P_{m,k}(x) = 1, \ x \in [0, 1],$$

(ii) Positivity :

$$\forall k \in \{0, \ldots, m\} \quad P_{m,k}(x) \geq 0,$$

(iii) Symmetry :

$$\forall k \in \{0, \ldots, m\} \quad P_{m,k}(x) = P_{m,m-k}(1-x),$$

(iv) recurrence formula: for $m > 0$,

$$P_{m,k}(x) = \begin{cases} (1-x)P_{m-1,k}(x) & si \;\; k = 0 \\ (1-x)P_{m-1,k}(x) + xP_{m-1,k-1}(x) & \forall k \in \{1, \ldots, m-1\} \\ xP_{m-1,k-1}(x) & si \;\; k = m. \end{cases}$$

Based on the definition above, Weierstrass Theorem can be restated as

Theorem 3.2. *Let $f : [0,1] \to \mathbb{R}$ be a continuous real-functions. The Bernstein polynomials of order m associate to f are give by :*

$$\forall m \in \mathbb{N},\ \forall x \in [0,1],\ B_m(f)(x) = \sum_{k=0}^{m} f\left(\frac{k}{m}\right) \binom{m}{k} x^k (1-x)^{m-k}.$$

Then, we have

$$\lim_{m\to\infty} \|f - B_m(f)\|_\infty = \lim_{m\to\infty} \sup_{x\in[0,1]} |f(x) - B_m(f)(x)| = 0.$$

In particular, any continuous function on $[0,1]$ is the uniform limit of a sequence of Bernstein polynomials.

Proof of Theorem 3.2. We shall compute the value of

$$\begin{aligned} T &= \sum_{k=0}^{m} (k - mx)^2 P_{m,k}(x) \\ &= \sum_{k=0}^{m} (k - mx)^2 \frac{m!}{k!(m-k)!} x^k (1-x)^{m-k} \\ &= \left[\sum_{k=0}^{m} k^2 \frac{m!}{k!(m-k)!} x^k (1-x)^{m-k} + \sum_{k=0}^{m} m^2 x^2 \frac{m!}{k!(m-k)!} x^k (1-x)^{m-k} \right. \end{aligned}$$

$$
\begin{aligned}
&\quad - \sum_{k=0}^{m} 2kmx \frac{m!}{k!(m-k)!} x^k (1-x)^{m-k} \Bigg] \\
&= mx \Bigg[\sum_{k=0}^{m} k \frac{(m-1)!}{(k-1)!(m-k)!} x^{k-1} (1-x)^{m-k} + mx \\
&\quad - 2mx \sum_{k=0}^{m} \frac{(m-1)!}{(k-1)!(m-k)!} x^{k-1} (1-x)^{m-k} \Bigg] \\
&= mx \Bigg[\sum_{k=1}^{m} k \frac{(m-1)!}{(k-1)!(m-k)!} x^{k-1} (1-x)^{m-k} - mx \Bigg] \\
&= mx \Bigg[\sum_{\ell=0}^{m-1} (\ell+1) \frac{(m-1)!}{\ell!(m-1-\ell)!} x^{\ell} (1-x)^{m-1-\ell} - mx \Bigg] \\
&= mx \Bigg[\sum_{\ell=0}^{m-1} (m-1) x \frac{(m-2)!}{(\ell-1)!(m-1-\ell)!} x^{\ell-1} (1-x)^{m-1-\ell} + 1 - mx \Bigg] \\
&= mx[(m-1)x + 1 - mx] \\
&= mx(1-x),
\end{aligned}
$$

since $x(1-x) \leq 1/4$ on $[0,1]$, we obtain the inequality

$$
\sum_{\left|\frac{k}{m}-x\right| \geq \delta} P_{m,k}(x) \leq \frac{1}{\delta^2} \sum_{\left|\frac{k}{m}-x\right| \geq \delta} \left(\frac{k}{m} - x\right)^2 P_{m,k}(x) \leq \frac{1}{m^2\delta^2} T = \frac{x(1-x)}{m\delta^2} \leq \frac{1}{4m\delta^2},
$$

for $\left|\frac{k}{m} - x\right|^2 / \delta^2 \geq 1$. If now the function f is bounded, say $|f(u)| \leq M$ in $0 \leq u \leq 1$ and x a point of continuity for a given $\epsilon > 0$, we can find a $\delta > 0$ such that $|x - x'|, \delta$ implies that $|f(x) - f(x')| < \epsilon$. We denote the Bernstein polynomial by $B_m(x)$, then we have

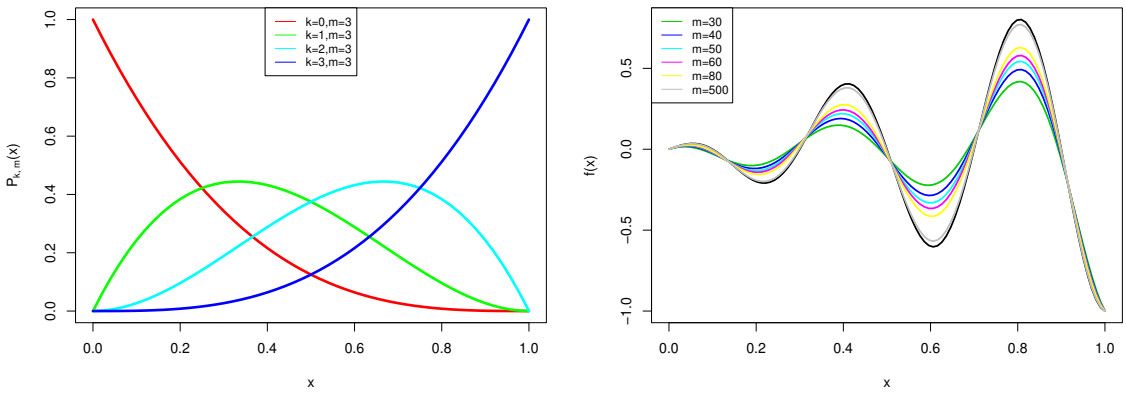
$$
|f(x) - B_m(x)| = \left| \sum_{k=0}^{m} \left\{ f(x) - f\left(\frac{k}{m}\right) \right\} P_{m,k}(x) \right|
$$

$$\leq \sum_{\left|\frac{k}{m}-x\right|<\delta} \left| f(x) - f\left(\frac{k}{m}\right) \right| P_{m,k}(x) + \sum_{\left|\frac{k}{m}-x\right|\geq\delta} \left| f(x) - f\left(\frac{k}{m}\right) \right| P_{m,k}(x)$$
$$\leq \epsilon \sum_{k=0}^{m} P_{m,k}(x) + 2M(4m\delta^2)^{-1}.$$

Therefore,

$$|f(x) - B_m(x)| \leq \epsilon + M(2m\delta^2)^{-1} \tag{3.1}$$

and if m si sufficiently large, $|f(x) - B_m(x)| < 2\epsilon$. Finally, if $f(x)$ is continuous in the whole interval $[0, 1]$ then (3.1) holds with a δ independent of x, so that $B_m(x) \to f(x)$ uniformly. This completes the proof. □



(a) (b)

Figure 3.1: (a) Bernstein polynomials, (b) Approximation of function $f(x) = x\cos(5\pi x)$ using Bernstein polynomials of degree $m = 30, 40, 50, 60, 80, 500$.

3.1.1 Bernstein distribution function estimator

Given a random sample $X_1, \ldots, X_n$ draw from a random variable X of distribution function F defined on $[0, 1]$. Motivated by the problem of smooth estimation of

F, Babu et al. (2002) proposed the univariate Bernstein estimator, which takes the following form

$$\widehat{F}_{n,m}(x) = \sum_{k=0}^{m} F_n\left(\frac{k}{m}\right) P_{m,k}(x), \qquad k = 0, ..., m. \tag{3.2}$$

where m is the smoothing parameter, $P_{m,k}(x) = \binom{m}{k} x^k (1-x)^{m-k}$ are binomial probabilities and F_n denotes the empirical distribution function constructed from a sample of size n. They have shown it to be uniformly strongly consistent when both n and m increase to infinity. This estimator was further studied by (Leblanc, 2009, 2012a,b) among other authors. The following theorem states the asymptotic properties of $\widehat{F}_{n,m}$

Theorem 3.3. *Assuming F is continuous (and bounded) and admits two continuous and bounded derivatives on $[0,1]$, we have for $x \in (0,1)$ that*

(i)

$$\mathbb{B}\text{ias}[\widehat{F}_{n,m}(x)] = \mathbb{E}\left[\widehat{F}_{n,m}(x)\right] - F(x) = m^{-1}b(x) + o(m^{-1}),$$

where $b(x) = 2^{-1}x(1-x)F''(x)$. Also, we have

(ii)

$$\mathbb{V}\text{ar}\left[\widehat{F}_{n,m}(x)\right] = n^{-1}\sigma^2(x) - n^{-1}m^{-1/2}V(x) + o(n^{-1}m^{-1/2}),$$

where $V(x) = f(x)[2x(1-x)/\pi]^{1/2}$ and $\sigma^2(x) = F(x)[1-F(x)]$.

(iii) And

$$\text{MSE}[\widehat{F}_{n,m}(x)] = n^{-1}\sigma^2(x) - n^{-1}m^{-1/2}V(x) + m^{-2}b^2(x) + o(m^{-2}) + o(n^{-1}m^{-1/2}),$$

as both $n, m \to 0$.

Proof of Theorem 3.3 is given in Appendix A.5. For a discussion of the asymptotic normality, see Babu et al. (2002, Theorem 3.2).

In the multivariate case, let $\mathbf{X} = (X_1, ..., X_d)$ denote a $d-$dimensional random vector, with a common cumulative distribution function F, with its associated density function f, supported on the $d-$dimensional hypercube. We assume for convenience (without loss of generality) that this support is the unit square $[0, 1]^d$. Obviously, it is possible to adapt our method to more general cases, when the data is defined on other intervals by taking appropriate transformations.

Babu and Chaubey (2006) introduced a Bernstein polynomial estimator for a distribution function F on a hypercube. Their Bernstein multivariate distribution function estimator is defined as follows

$$\widehat{F}_{m,n}(x_, \ldots, x_d) = \sum_{k_1=0}^{m} \cdots \sum_{k_d=0}^{m} F_n\left(\frac{k_1}{m}, \ldots, \frac{k_d}{m}\right) \prod_{j=1}^{d} P_{k_j,m}(x_j). \tag{3.3}$$

They have shown it to be uniformly strongly consistent when $n, m \to \infty$. Note that $\widehat{F}$ is a proper distribution function and a polynomial in x_j. Recently, Belalia (2016) derived the asymptotic bias, variance and normality of this estimator. He also identified the asymptotically optimal choice of the parameter m in the sense of MSE. Under the following notations:

1. F_X (resp.f_X) and F_Y (resp. f_Y) are the marginal distribution functions (resp. densities) of X and Y.

2. F_x, F_y, F_{xx}, F_{yy} and F_{xy} are the first and second partial order derivatives of F.

The asymptotic properties of $\hat{F}_{m,n}(x,y)$ can be stated in the following theorem from Belalia (2016).

Theorem 3.4. *Assume that F is continuous and all its partial derivatives up to the second order are continuous and bounded on $[0,1]^2$. We have for $x, y \in [0,1]$ that*

(i) $\mathbb{E}[\hat{F}_{m,n}(x,y)] = F_m(x,y)$

$$= \begin{cases} F(x,y) + m^{-1}B(x,y) + o(m^{-1}) & \text{if } 0 < x, y < 1 \\ 0 & \text{if } x = 0 \text{ and/ or } y = 0 \\ F_X(x) + m^{-1}b(x)f'_X(x) + o(m^{-1}) & \text{if } 0 < x < 1, y = 1 \\ F_Y(y) + m^{-1}b(y)f'_Y(y) + o(m^{-1}) & \text{if } x = 1, 0 < y < 1 \\ 1 & \text{if } (x,y) = (1,1)\ , \end{cases}$$

where $B(x,y)$ and $b(z)$ are defined by

$$B(x,y) = \frac{1}{2}[x(1-x)F_{xx}(x,y) + y(1-y)F_{yy}(x,y)], \quad b(z) = z(1-z)/2.$$

(ii) $\mathbb{V}ar[\hat{F}_{m,n}(x,y)]$

$$= \begin{cases} n^{-1}\sigma^2(x,y) - m^{-\frac{1}{2}}n^{-1}V(x,y) + o(m^{-\frac{1}{2}}n^{-1}) & \text{if } 0 < x, y < 1 \\ 0 & \text{if } x = 0 \text{ and/ or } y = 0 \\ n^{-1}\sigma^2(x) - m^{-\frac{1}{2}}n^{-1}V_X(x) + o(m^{-\frac{1}{2}}n^{-1}) & \text{if } 0 < x < 1 \text{ and } y = 1 \\ n^{-1}\sigma^2(y) - m^{-\frac{1}{2}}n^{-1}V_Y(y) + o(m^{-\frac{1}{2}}n^{-1}) & \text{if } x = 1 \text{ and } 0 < y < 1 \\ 0 & \text{if } (x,y) = (1,1)\ , \end{cases}$$

where,

$$V(x,y) = \left\{ F_x(x,y) \left(2x(1-x)/\pi\right)^{1/2} + F_y(x,y) \left(2y(1-y)/\pi\right)^{1/2} \right\},$$

$$\sigma^2(x,y) = F(x,y)[1 - F(x,y)],$$

and for $Z = X$ *or* Y*,*

$$\sigma_Z^2(z) = F_Z(z)[1 - F_Z(z)],\ V_Z(z) = \left\{ F_Z(z) \left(2z(1-z)/\pi\right)^{1/2} \right\}.$$

The proof of this theorem can be found in Belalia (2016). □

3.1.2 Bernstein probability density function estimator

Assume that the cumulative distribution F has an associated density f. Suppose that f is continuous (and bounded) and admits two continuous and bounded derivatives on $[0,1]$. Through differentiation, the estimator (3.2) naturally leads to a density estimator

$$\begin{aligned}
\widehat{f}_{n,m}(x) = \frac{d}{dx}\widehat{F}_{n,m}(x) &= \frac{d}{dx}\sum_{k=0}^{m} F_n\left(\frac{k}{m}\right) P_{m,k}(x) \\
&= \sum_{k=0}^{m} F_n\left(\frac{k}{m}\right) \frac{d}{dx}\left[\frac{m!}{k!(m-k)!}x^k(1-x)^{m-k}\right] \\
&= \sum_{k=0}^{m} F_n\left(\frac{k}{m}\right) \frac{m!}{k!(m-k)!}\left[kx^{k-1}(1-x)^{m-k} - x^k(1-x)^{m-k-1}(m-k)\right] \\
&= \sum_{k=0}^{m} F_n\left(\frac{k}{m}\right) \frac{m!}{k!(m-k)!}kx^{k-1}(1-x)^{m-k} \\
&\quad - \sum_{k=0}^{m} F_n\left(\frac{k}{m}\right) \frac{m!}{k!(m-k)!}x^k(1-x)^{m-k-1}(m-k)
\end{aligned}$$

$$= \sum_{k=1}^{m} F_n\left(\frac{k}{m}\right) \frac{m!}{(k-1)!(m-k)!} x^{k-1}(1-x)^{m-k} - \sum_{k=0}^{m} F_n\left(\frac{k}{m}\right) \frac{m!}{k!(m-k)!} x^{k}(1-x)^{m-k-1}(m-k).$$

Put $\ell = k - 1$,

$$\begin{aligned}\widehat{f}_{n,m}(x) &= m\left[\sum_{\ell=0}^{m-1} F_n\left(\frac{\ell+1}{m}\right) \frac{(m-1)!}{\ell!(m-\ell-1)!} x^{\ell}(1-x)^{m-\ell-1} - \sum_{k=0}^{m-1} F_n\left(\frac{k}{m}\right) \frac{(m-1)!}{k!(m-k-1)!} x^{k}(1-x)^{m-k-1}\right] \\ &= m \sum_{k=0}^{m-1} \left[F_n\left(\frac{k+1}{m}\right) - F_n\left(\frac{k}{m}\right)\right] P_{m-1,k}(x), \end{aligned} \tag{3.4}$$

is a polynomial of degree $(m-1)$. This estimator can be further written as a finite mixture of Beta densities with data-driven weights:

$$\widehat{f}_{m,n}(x) = \sum_{k=0}^{m-1} W_{k,m}\, \beta_{k+1,m-k}(x), \tag{3.5}$$

where $W_{k,m} = F_n\left((k+1)/m\right) - F_n(k/m)$ form a sequence of nonnegative weights that sum to unity and $\beta_{a,b}$ stands for the beta density with parameters $a, b > 0$. From this, we see that $\widehat{f}_{m,n}$ is a density for any observed sample. We note that this estimator can also be represented as

$$\widehat{f}_{m,n}(x) = \frac{m}{n} \sum_{k=0}^{m-1} M_{k,m}\, P_{m-1,k}(x), \tag{3.6}$$

where $M_{k,m}$ corresponds to the number of observations falling in the interval

$$A_{k,m} = \left(\frac{k}{m}, \frac{k+1}{m}\right], \qquad \text{for } k = 0, 1, \ldots, m-1.$$

In other words, $M_{0,m}, M_{1,m}, \ldots, M_{m-1,m}$ correspond to the bin counts obtained from a histogram constructed with m bins of equal length over the unit interval. The Bernstein density estimator was originally introduced by Vitale (1975), who has shown it to be consistent in the Mean Squared Error (MSE) when $m \to \infty$ and $mn^{-1} \to 0$ as $n \to \infty$.

The asymptotic properties of estimator (3.4) are stated in the following theorem

Theorem 3.5. *Assuming f is continuous (and bounded) and admits two continuous and bounded derivatives on $[0,1]$, we have for $x \in (0,1)$ that*

$$\mathbb{E}\left[\widehat{f}_{n,m}(x)\right] = f(x) + m^{-1}\Delta_1(x) + m^{-2}1/6[1-6x(1-x)]f''(x) + E_{B,f,m}(x),$$

where $\Delta_1(x) = 1/2[(1-2x)f'(x) + x(1-x)f''(x)]$, and $E_{B,f,m}(x) = o(T_{2,m}(x)) + o(m^{-1}[T_{2,m}(x) + 1/m^2]^{1/2}) + o(m^{-2})$. Also we have

$$\mathbb{V}\text{ar}\left[\widehat{f}_{n,m}(x)\right] = \frac{m}{n} f(x) S_{m-1}(x) + E_{V,f,m}(x),$$

where $E_{V,f,m}(x) = O(mn^{-1}[T_{2,m-1}(x)S_{m-1}(x)]^{1/2}) + O(n^{-1})$. And

$$\text{MSE}[\widehat{f}_{n,m}(x)] = n^{-1}m^{1/2}f(x)\psi_1(x) + m^{-2}\Delta_1^2(x) + o(n^{-1}m^{1/2}) + o(m^{-2}),$$

where $\psi_1(x)$ is defined as in Lemma 3.

Turning our attention to the multivariate case, following Babu and Chaubey (2006), the Bernstein estimator of order m of the joint cumulative distribution function F is defined by (3.3). For the sake of clarity, we consider here the bivariate case. Applying a second order mixed derivative, this estimator naturally leads to a smooth estimator of f. Specifically, we have

$$\begin{aligned}\hat{f}_{m,n}(x,y) &= \frac{\partial^2}{\partial x \partial y}\hat{F}_{m,n}(x,y) \\ &= \sum_{k=0}^{m}\sum_{\ell=0}^{m} F_n\left(\frac{k}{m},\frac{\ell}{m}\right)\frac{d}{dx}P_{m,k}(x)\,\frac{d}{dy}P_{m,\ell}(y) \\ &= m^2\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} B_{k,\ell,m}^{(n)}\,P_{m-1,k}(x)\,P_{m-1,\ell}(y),\end{aligned}$$

where

$$B_{k,\ell,m}^{(n)} = F_n\Big(\frac{k+1}{m},\frac{\ell+1}{m}\Big) - F_n\Big(\frac{k+1}{m},\frac{\ell}{m}\Big) - F_n\Big(\frac{k}{m},\frac{\ell+1}{m}\Big) + F_n\Big(\frac{k}{m},\frac{\ell}{m}\Big),$$

and F_n denotes the bivariate empirical distribution constructed from a sample of size n.

Now, let $M_{k,\ell,m}$ denote the numbers of pairs (X_i, Y_i) inside the square

$$A_{k,\ell,m} = \left\{(s,t) : \frac{k}{m} < s \le \frac{k+1}{m}, \frac{\ell}{m} < t \le \frac{\ell+1}{m}\right\},$$

for $k, \ell = 0, 1, \ldots, m-1$. Then, by observing that $B_{k,\ell,m}^{(n)} = \frac{1}{n}M_{k,\ell,m}$, we can rewrite $\hat{f}_{m,n}$ as

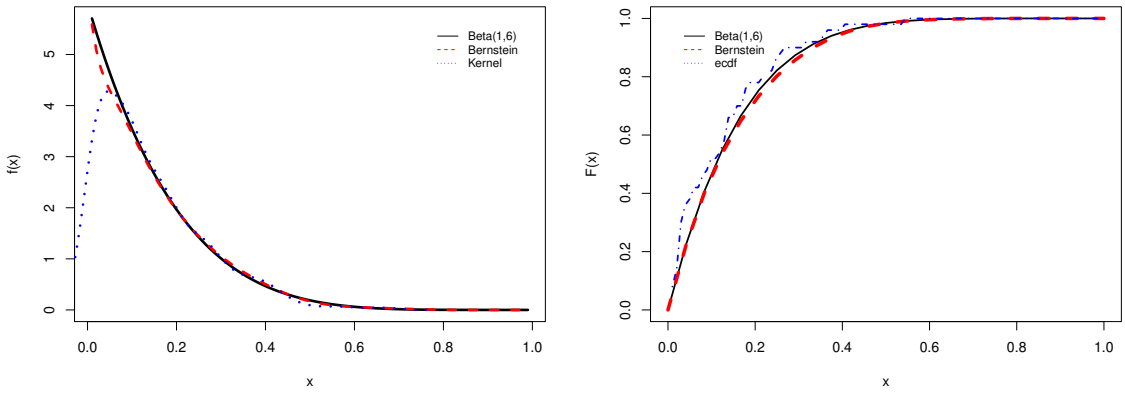
$$\hat{f}_{m,n}(x,y) = \frac{m^2}{n}\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} M_{k,\ell,m}\,P_{m-1,k}(x)\,P_{m-1,\ell}(y). \tag{3.7}$$

This is the original expression for the Bernstein estimator of a bivariate density defined

on the unit square as it was proposed by Tenbusch (1994).

3.1.3 Numerical Illustration

To illustrate the effectiveness of Bernstein distribution estimators (3.2) and (3.4) the Beta$(1, 6)$ cumulative distribution function and its associated density are used. Figure 3.2a shows the Bernstein density estimator (3.4) of degree $m = 50$ (red dashed line) compared to the kernel estimator (blue dotted line) with bandwidth parameter $h = 0.0302$. We point out that the Bernstein estimator has a good performance, in particular close to the boundary $x = 0$. The Bernstein cumulative distribution function estimator with degree $m = 35$ (red dashed line), and the empirical distribution function (blue dotted line) are depicted in Figure 3.2b. Also, one can notice the smoothness of Bernstein estimator against the empirical distribution function.



(a) (b)

Figure 3.2: (a) Bernstein density estimator compared to kernel estimator, (b) Bernstein cumulative distribution compared to the empirical distribution.

3.2 Bernstein conditional density estimation

For $x \in [0,1]$ such that $g(x) > 0$, the conditional density of Y given $X = x$ is given by

$$f_x(y) = \frac{f(x,y)}{g(x)} \qquad \text{for } y \in [0,1],$$

and hence, can be simply viewed as the ratio of two unconditional densities. This leads to a simple strategy for the estimation of f_x. Indeed, an estimator of f_x is naturally defined through

$$\hat{f}_x(y) = \frac{\hat{f}(x,y)}{\hat{g}(x)}, \tag{3.8}$$

where $\hat{f}$ and $\hat{g}$ are consistent estimators of the joint density f and of the marginal density g, respectively. This approach was first used in a context of kernel estimation by Rosenblatt (1969), and has been used by many other authors since then (e.g. Hyndman et al., 1996; Bashtannyk and Hyndman, 2001; Hall et al., 2004); see the interesting discussion presented by Efromovich (2007).

Recently, Belalia et al. (2019) proposed a new estimator for f_x based on Bernstein polynomials. At this point, our new estimator of the conditional density function f_x can be defined via (3.8) as

$$\hat{f}_{x,m,n}(y) = \frac{\hat{f}_{m,n}(x,y)}{\hat{g}_{m,n}(x)}, \tag{3.9}$$

where $\hat{g}_{m,n}$ and $\hat{f}_{m,n}$ are respectively defined in (3.6) and (3.7). We refer to this estimator as the Bernstein estimator of order m of the conditional density f_x. The proposed estimator is clearly nonnegative and is a genuine conditional density for any value of x. To see this, one can see that it can be written as a mixture of beta

densities with data-driven weights. Specifically, we have that

$$\hat{f}_{x,m,n}(y) = \sum_{\ell=0}^{m-1} W_{x,\ell,m}\, \beta_{\ell+1,m-\ell}(y), \tag{3.10}$$

where

$$W_{x,\ell,m} = \frac{\sum_{k=0}^{m-1} M_{k,\ell,m}\, P_{m-1,k}(x)}{\sum_{k=0}^{m-1} M_{k,m}\, P_{m-1,k}(x)}.$$

The weights $W_{x,\ell,m}$ are nonnegative and sum to unity since

$$\sum_{\ell=0}^{m-1} M_{k,\ell,m} = M_{k,m}.$$

Belalia et al. (2019) studied the asymptotic properties of $\hat{f}_{m,n}$, which include the asymptotic bias, variance, and distribution limit.

Chapter 4

Two-Stage Conditional Density Estimation Based on Bernstein Polynomials

In the previous chapters two main nonparametric estimation methods were discussed, namely, kernel based estimation methods, and nonparametric estimation methods based on Bernstein polynomials. In this chapter a conditional density estimator is presented and studied, the proposed approach will combine the previous kernel and Bernstein based methods.

4.1 Two-Stage Conditional Density Estimator

Recently, Belalia et al. (2017) have combined both methods to construct a two-stage estimator of the conditional distribution function F_x, their estimator is defined as

follows

$$\widehat{F}_{x,mh}(y) = B_{\hat{F}_{x,h,m}}(y) = \sum_{k=0}^{m} \hat{F}_{x,h}(k/m)\, P_{m,k}(y), \tag{4.1}$$

where $\hat{F}_{x,h}$ is the Nadaraya-Watson estimator of F_x defined as in Chapter 2 by

$$\hat{F}_{x,h}(y) = \frac{\sum_{i=1}^{n} K_h(x - X_i)\, \mathbb{I}(Y_i \leq y)}{\sum_{j=1}^{n} K_h(x - X_j)} = \sum_{i=1}^{n} w_i(x,h)\, \mathbb{I}(Y_i \leq y), \tag{4.2}$$

where the weights $w_i = K_h(x - X_i)/\sum_{j=1}^{n} K_h(x - X_j)$, $K_h(x) = h^{-1}K(x/h)$ with K is a kernel function and h is bandwidth parameter. Typically, K is taken to be an symmetric density function and $h = h_n$ is a deterministic sequence depending on n in such a way that $h_n \to 0$ as $n \to \infty$.

It was shown in Belalia et al. (2017) that the estimator (4.1) comes with a companion density estimators. Indeed, differentiation with respect to y leads to the following simple estimator of f_x,

$$\widehat{f}_{x,mh}(y) = \frac{d}{\mathrm{d}y}\widehat{F}_{x,mh}(y) = m \sum_{k=0}^{m-1} \Big[\hat{F}_{x,h}\Big([k+1]/m\Big) - \hat{F}_{x,h}(k/m)\Big]\, P_{m-1,k}(y), \tag{4.3}$$

which is a polynomial of degree $m - 1$. We point out that this estimator can be rewritten as data driven mixture Beta densities, specifically, we have

$$\widehat{f}_{x,mh}(y) = \sum_{k=0}^{m-1} W_{m,k}\beta_{k+1,m-k}(y), \tag{4.4}$$

where $W_{m,k} = \Big[\hat{F}_{x,h}\Big([k+1]/m\Big) - \hat{F}_{x,h}(k/m)\Big]$ form a sequence of non-negative weights and $\beta_{a,b}$ stands for the beta density with parameters $a, b > 0$.

This chapter is devoted to the study of the asymptotic behaviour of the Bernstein

conditional density estimator, including its asymptotic bias, variance and integrated mean squared error (IMSE). We also establish its asymptotic normality. Our results are based on the following regularity conditions.

Assumption 1. The marginal density of X, denoted $g(x)$, is twice continuously differentiable with respect to x, with bounded second derivative. The conditional distribution function $F_x(y)$ is twice continuously differentiable with respect to both x and y, the first and second order derivatives being bounded.

Assumption 2. The kernel function K is a symmetric, bounded and compactly supported density function.

Assumption 3. As $n \to \infty$, we also have $h \to 0$, $nh \to \infty$ and $m \to \infty$.

In what follows, we use the following notation

$$F_x^{(i,j)}(y) = \frac{\partial^{i+j}}{\partial x^i \partial y^j} F_x(y), \quad i,j = 0,1,2,$$
$$\kappa_2 = \int_{\mathbb{R}} y^2 K(y)\, dy, \qquad \kappa = \int_{\mathbb{R}} K^2(y)\, dy.$$

We point out that Assumption 2 implies both κ and κ_2 are finite.
Before stating our main results, some needed auxiliary intermediate results are provided in the following lemma.

Lemma 1. *Under Assumption 1, we have*

1. $\sum_{k=0}^{m-1} F_x^{(0,1)}\left(k/m\right) P_{m-1,k}(y) = F_x^{(0,1)}(y) - m^{-1} y F_x^{(0,2)}(y) + o(m^{-1})$,

2. $\sum_{k=0}^{m-1} F_x^{(0,2)}\left(k/m\right) P_{m-1,k}(y) = F_x^{(0,2)}(y) + o(1)$,

3. $\sum_{k=0}^{m-1} F_x^{(1,1)}(k/m)\, P_{m-1,k}(y) = F_x^{(1,1)}(y) + o(1)$,

4. $\sum_{k=0}^{m-1} F_x^{(2,1)}(k/m)\, P_{m-1,k}(y) = F_x^{(2,1)}(y) + o(1)$.

Proof of Lemma 1. Using Taylor expansion, we get

$$F_x^{(0,1)}(k/m) = F_x^{(0,1)}(y) + (k/m - y)\, F_x^{(0,2)}(y) + o(k/m - y). \tag{4.5}$$

This expansion along with the fact that

$$\sum_{k=0}^{m-1} (k/m - y)\, P_{m-1,k}(y) = [(m-1)T_{m-1,1}(y) - yT_{m-1,0}(y)]/m = -y/m,$$

where $T_{m-1,j}(y) = (m-1)^{-j} \sum_{k=0}^{m-1} (k - (m-1)y)^j P_{m-1,k}(y)$ for $j = 0, 1$.
Substituting the result into equation (4.5), and doing the same expansion for $F_x^{(0,2)}(k/m)$, $F_x^{(1,1)}(k/m)$ and $F_x^{(2,1)}(k/m)$, we obtain the Lemma 1. □

4.2 Asymptotic Bias

To provide the asymptotic bias of our estimator (4.3), we first state an intermediate result that calculates the asymptotic expectation of

$$\widehat{N}_x(y) = \left[\widehat{f}_{x,mh}(y) - f_x(y)\right]\hat{G}(x),$$

where $\hat{G}(x)$ corresponds, up to a factor $1/n$, to the denominator in the expression of the estimator (4.3), that is

$$\hat{G}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i).$$

The following proposition provides the asymptotic expectation of $\widehat{N}_x(y)$, which will be used to establish the asymptotic bias of the proposed estimator.

Proposition 2. *Under Assumption 1 to 3, we have*

$$\mathbb{E}\left(\widehat{N}_x(y)\right) = g(x)\Big[- m^{-1} y f_x'(y) + \frac{1}{2} m^{-1} f_x'(y) + h^2 \kappa_2 \frac{g'(x)}{g(x)} F_x^{(1,1)}(y) + \frac{h^2 \kappa_2}{2} F_x^{(2,1)}(y)\Big] + o(h^2) + o(m^{-1}). \tag{4.6}$$

Proof of Proposition 2. First, we rewrite $\widehat{N}_x(y)$, as a sum of independent random variables

$$\begin{aligned}
\widehat{N}_x(y) &= \left\{ m \sum_{k=0}^{m-1} \left[\hat{F}_{x,h}\Big([k+1]/m\Big) - \hat{F}_{x,h}(k/m) - m^{-1} f_x(y) \right] P_{m-1,k}(y) \right\} \hat{G}(x) \\
&= m \sum_{k=0}^{m-1} \left[\frac{\sum_{i=1}^n K_h(x - X_i)\, \mathbb{I}(Y_i \le \frac{k+1}{m})}{\sum_{j=1}^n K_h(x - X_j)} - \frac{\sum_{i=1}^n K_h(x - X_i)\, \mathbb{I}(Y_i \le \frac{k}{m})}{\sum_{j=1}^n K_h(x - X_j)} \right] P_{m-1,k}(y) \hat{G}(x) \\
&\quad - \sum_{k=0}^{m-1} f_x(y) P_{m-1,k}(y) \hat{G}(x) \\
&= \frac{m}{n} \sum_{i=1}^n \sum_{k=0}^{m-1} \left[\mathbb{I}\left(Y_i \le \frac{k+1}{m}\right) - \mathbb{I}\left(Y_i \le \frac{k}{m}\right) - m^{-1} f_x(y) \right] P_{m-1,k}(y) K_h(x - X_i) \\
&= \frac{m}{n} \sum_{i=1}^n Z_{i,m},
\end{aligned} \tag{4.7}$$

where

$$Z_{i,m} = \sum_{k=0}^{m-1} \left[\mathbb{I}\left(Y_i \le \frac{k+1}{m}\right) - \mathbb{I}\left(Y_i \le \frac{k}{m}\right) - m^{-1} f_x(y) \right] P_{m-1,k}(y) K_h(x - X_i).$$

Since for a given m the random variables $Z_{1,m}, ..., Z_{n,m}$ are *i.i.d.*, one can write that

$$\begin{aligned}
\mathbb{E}[\widehat{N}_x(y)] &= m\mathbb{E}\left[Z_{1,m}\right] \\
&= m\sum_{k=0}^{m-1}\mathbb{E}\left\{\left[\mathbb{I}\left(Y_1 \leq \frac{k+1}{m}\right) - \mathbb{I}\left(Y_1 \leq \frac{k}{m}\right) - m^{-1}f_x(y)\right]K_h(x-X_1)\right\}P_{m-1,k}(y) \\
&= m\sum_{k=0}^{m-1}\mathbb{E}\left[\mathbb{I}\left(Y_1 \leq \frac{k+1}{m}\right)K_h(x-X_1)\right]P_{m-1,k}(y) \\
&\quad - m\sum_{k=0}^{m-1}\mathbb{E}\left[\mathbb{I}\left(Y_1 \leq \frac{k}{m}\right)K_h(x-X_1)\right]P_{m-1,k}(y) - f_x(y)\mathbb{E}\left[K_h(x-X_1)\right].
\end{aligned} \tag{4.8}$$

From the same calculation as Equation (17) in Belalia et al. (2017), we can have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{I}\left(Y_1 \leq \frac{k}{m}\right)K_h(x-X_1)\right] &= g(x)F_x\left(\frac{k}{m}\right) + \frac{h^2\kappa_2}{2}\left[g(x)F_x^{(2,0)}\left(\frac{k}{m}\right)\right. \\
&\quad \left. +2g'(x)F_x^{(1,0)}\left(\frac{k}{m}\right) + g''(x)F_x\left(\frac{k}{m}\right)\right] + o\left(h^2\gamma\left(\frac{k}{m}\right)\right),
\end{aligned}$$

where $\gamma(\cdot)$ is a function on the support $[0,1]$. Similarly, we get

$$\begin{aligned}
\mathbb{E}\left[\mathbb{I}\left(Y_1 \leq \frac{k+1}{m}\right)K_h(x-X_1)\right] &= g(x)F_x\left(\frac{k+1}{m}\right) + \frac{h^2\kappa_2}{2}\left[g(x)F_x^{(2,0)}\left(\frac{k+1}{m}\right)\right. \\
&\quad \left. +2g'(x)F_x^{(1,0)}\left(\frac{k+1}{m}\right) + g''(x)F_x\left(\frac{k+1}{m}\right)\right] \\
&\quad + o\left(h^2\gamma\left(\frac{k+1}{m}\right)\right).
\end{aligned}$$

Thus, equation (4.8) is equivalent to

$$\begin{aligned}\mathbb{E}\left[\widehat{N}_x(y)\right] = m\sum_{k=0}^{m-1}\Bigg\{ & g(x)\left[F_x\left(\frac{k+1}{m}\right) - F_x\left(\frac{k}{m}\right)\right] + \frac{h^2\kappa_2}{2}g(x)\left[F_x^{(2,0)}\left(\frac{k+1}{m}\right) - F_x^{(2,0)}\left(\frac{k}{m}\right)\right] \\ & + h^2\kappa_2 g'(x)\left[F_x^{(1,0)}\left(\frac{k+1}{m}\right) - F_x^{(1,0)}\left(\frac{k}{m}\right)\right] + \frac{h^2\kappa_2}{2}g''(x)\left[F_x\left(\frac{k+1}{m}\right) - F_x\left(\frac{k}{m}\right)\right] \\ & + \left[o\left(h^2\gamma\left(\frac{k+1}{m}\right)\right) - o\left(h^2\gamma\left(\frac{k}{m}\right)\right)\right]\Bigg\}P_{m-1,k}(y) - f_x(y)\mathbb{E}[K_h(x-X_1)].\end{aligned} \tag{4.9}$$

By using Taylor expansion for $F_x\left(\frac{k+1}{m}\right)$, $F_x^{(1,0)}\left(\frac{k+1}{m}\right)$ and $F_x^{(2,0)}\left(\frac{k+1}{m}\right)$ around $\frac{k}{m}$ we get

$$\begin{aligned}F_x\left(\frac{k+1}{m}\right) &= F_x\left(\frac{k}{m}\right) + m^{-1}f_x\left(\frac{k}{m}\right) + \frac{1}{2}m^{-2}f_x'\left(\frac{k}{m}\right) + o(m^{-2}), \\ F_x^{(1,0)}\left(\frac{k+1}{m}\right) &= F_x^{(1,0)}\left(\frac{k}{m}\right) + m^{-1}F_x^{(1,1)}\left(\frac{k}{m}\right) + o(m^{-1}), \\ F_x^{(2,0)}\left(\frac{k+1}{m}\right) &= F_x^{(2,0)}\left(\frac{k}{m}\right) + m^{-1}F_x^{(2,1)}\left(\frac{k}{m}\right) + o(m^{-1}).\end{aligned}$$

By taking the same expansion for $\gamma\left(\frac{k+1}{m}\right)$, and then substituting back in the equation (4.9), we get

$$\begin{aligned}\mathbb{E}\left[\widehat{N}_x(y)\right] = m\sum_{k=0}^{m-1}\Bigg\{ & g(x)\left[f_x\left(\frac{k}{m}\right)m^{-1} + \frac{1}{2}f_x'\left(\frac{k}{m}\right)m^{-2} + o(m^{-2})\right] \\ & + \frac{h^2\kappa_2}{2}g(x)\left[F_x^{(2,1)}\left(\frac{k}{m}\right)m^{-1} + o(m^{-1})\right] + h^2\kappa_2 g'(x)\left[F_x^{(1,1)}\left(\frac{k}{m}\right)m^{-1} + o(m^{-1})\right]\end{aligned}$$

$$+ \frac{h^2\kappa_2}{2} g''(x) \left[f_x\left(\frac{k}{m}\right) m^{-1} + \frac{1}{2} f'_x\left(\frac{k}{m}\right) m^{-2} + o(m^{-2}) \right] + o(m^{-1}h^2) \Bigg\} P_{m-1,k}(y)$$

$$- f_x(y)\mathbb{E}[K_h(x - X_1)].$$

By the Equation (16) in Belalia et al. (2017), we have

$$\mathbb{E}[K_h(x - X_1)] = g(x) + \frac{h^2}{2}\kappa_2 g''(x) + o(h^2),$$

then according to Lemma 1 we can simplify the result as

$$\begin{aligned}
\mathbb{E}\left[\widehat{N}_x(y)\right] &= m\left[g(x) + \frac{h^2\kappa_2}{2} g''(x)\right] \sum_{k=0}^{m-1} f_x\left(\frac{k}{m}\right) m^{-1} P_{m-1,k}(y) \\
&\quad + m\left[g(x)o(m^{-2}) + \frac{h^2\kappa_2}{2} g(x)o(m^{-1}) + h^2\kappa_2 g'(x)o(m^{-1}) + \frac{h^2\kappa_2}{2} g''(x)o(m^{-2})\right] \\
&\quad + m\left[\frac{1}{2} g(x)m^{-2} + \frac{h^2\kappa_2}{4} g''(x)m^{-2}\right] \sum_{k=0}^{m-1} f'_x\left(\frac{k}{m}\right) P_{m-1,k}(y) \\
&\quad + h^2\kappa_2 g'(x) \sum_{k=0}^{m-1} F_x^{(1,1)}\left(\frac{k}{m}\right) P_{m-1,k}(y) + \frac{h^2\kappa_2}{2} g(x) \sum_{k=0}^{m-1} F_x^{(2,1)}\left(\frac{k}{m}\right) P_{m-1,k}(y) \\
&\quad + o(h^2) - f_x(y)\left[g(x) + \frac{h^2}{2}\kappa_2 g''(x) + o(h^2)\right] \\
&= \left[g(x) + \frac{h^2\kappa_2}{2} g''(x)\right]\left[f_x(y) - m^{-1} y f'_x(y) + o(m^{-1})\right] \\
&\quad + \left[g(x)o(m^{-1}) + \frac{h^2\kappa_2}{2} g(x)o(1) + h^2\kappa_2 g'(x)o(1) + \frac{h^2\kappa_2}{2} g''(x)o(m^{-1})\right]
\end{aligned}$$

$$+\left[\frac{1}{2}g(x)m^{-1}+\frac{h^2\kappa_2}{4}g''(x)m^{-1}\right]\left[f'_x(y)+o(1)\right]+h^2\kappa_2 g'(x)\left[F_x^{(1,1)}(y)+o(1)\right]$$

$$+\frac{h^2\kappa_2}{2}g(x)\left[F_x^{(2,1)}(y)+o(1)\right]+o(h^2)-f_x(y)\left[g(x)+\frac{h^2}{2}\kappa_2 g''(x)+o(h^2)\right]$$

$$=g(x)f_x(y)-g(x)m^{-1}yf'_x(y)+g(x)o(m^{-1})+\frac{h^2\kappa_2}{2}g''(x)f_x(y)-\frac{h^2\kappa_2}{2}g''(x)m^{-1}yf'_x(y)$$

$$+\frac{h^2\kappa_2}{2}g''(x)o(m^{-1})+\left[g(x)o(m^{-1})+\frac{h^2\kappa_2}{2}g(x)o(1)+h^2\kappa_2 g'(x)o(1)\right.$$

$$\left.+\frac{h^2\kappa_2}{2}g''(x)o(m^{-1})\right]+\frac{1}{2}g(x)m^{-1}f'_x(y)+\frac{h^2\kappa_2}{4}g''(x)m^{-1}f'_x(y)+\frac{1}{2}g(x)m^{-1}o(1)$$

$$+\frac{h^2\kappa_2}{4}g''(x)m^{-1}o(1)+h^2\kappa_2 g'(x)F_x^{(1,1)}(y)+h^2\kappa_2 g'(x)o(1)+\frac{h^2\kappa_2}{2}g(x)F_x^{(2,1)}(y)$$

$$+\frac{h^2\kappa_2}{2}g(x)o(1)+o(h^2)-g(x)f_x(y)-\frac{h^2\kappa_2}{2}g''(x)f_x(y)-f_x(y)o(h^2)$$

$$=g(x)\Big[-m^{-1}yf'_x(y)-\frac{h^2\kappa_2}{2}\frac{g''(x)}{g(x)}m^{-1}yf'_x(y)+\frac{1}{2}m^{-1}f'_x(y)+\frac{h^2\kappa_2}{4}\frac{g''(x)}{g(x)}m^{-1}f'_x(y)$$

$$+h^2\kappa_2\frac{g'(x)}{g(x)}F_x^{(1,1)}(y)+\frac{h^2\kappa_2}{2}F_x^{(2,1)}(y)\Big]+g(x)o(m^{-1})+\frac{h^2\kappa_2}{2}g''(x)o(m^{-1})$$

$$+\left[g(x)o(m^{-1})+\frac{h^2\kappa_2}{2}g(x)o(1)+h^2\kappa_2 g'(x)o(1)+\frac{h^2\kappa_2}{2}g''(x)o(m^{-1})\right]$$

$$+\frac{1}{2}g(x)m^{-1}o(1)+\frac{h^2\kappa_2}{4}g''(x)m^{-1}o(1)+h^2\kappa_2 g'(x)o(1)+\frac{h^2\kappa_2}{2}g(x)o(1)$$

$$+o(h^2)-f_x(y)o(h^2)$$

$$= g(x)\Big[- m^{-1} y f_x'(y) - \frac{h^2\kappa_2}{2}\frac{g''(x)}{g(x)} m^{-1} y f_x'(y) + \frac{1}{2} m^{-1} f_x'(y) + \frac{h^2\kappa_2}{4}\frac{g''(x)}{g(x)} m^{-1} f_x'(y)$$

$$+ h^2\kappa_2 \frac{g'(x)}{g(x)} F_x^{(1,1)}(y) + \frac{h^2\kappa_2}{2} F_x^{(2,1)}(y)\Big] + o(h^2) + o(m^{-1})$$

$$= g(x)\Big[- m^{-1} y f_x'(y) + \frac{1}{2} m^{-1} f_x'(y) + h^2\kappa_2 \frac{g'(x)}{g(x)} F_x^{(1,1)}(y) + \frac{h^2\kappa_2}{2} F_x^{(2,1)}(y)\Big]$$

$$+ o(h^2) + o(m^{-1}),$$

which completes the proof. □

Employing the same strategy as in Li and Racine (2007, Section 6.1), one can rewrite the difference between the density function and its estimate as

$$\widehat{f}_{x,mh}(y) - f_x(y) = \frac{\widehat{N}_x(y)}{g(x)} + O_p\left(h^2 + (nh)^{-1/2}\right), \tag{4.10}$$

the asymptotic bias (denoted as $\mathbb{AB}$ias) of the two-stage estimator $\widehat{f}_{x,mh}(y)$ can be deduced, and is given in the following theorem.

Theorem 4.1. *Under Assumptions 1-3, we have for $y \in (0,1)$ that*

$$\mathbb{AB}ias\left[\widehat{f}_{x,mh}(y)\right] = \left[(2m)^{-1} - m^{-1}y\right] f_x'(y) + h^2\kappa_2 \frac{g'(x)}{g(x)} F_x^{(1,1)}(y) + \frac{h^2\kappa_2}{2} F_x^{(2,1)}(y). \tag{4.11}$$

Note as m tends to infinity, and h to 0, $\widehat{f}_{x,mh}(y)$ becomes asymptotically unbiased.

4.3 Asymptotic Variance

In the goal to provide the asymptotic variance of our estimator $\widehat{f}_{x,mh}(y)$, we begin by calculating the asymptotic variance of $\widehat{N}_x(y)$ by stating the following lemma.

Proposition 3. *Under Assumption* $1-3$*, we have for* $y \in (0,1)$ *that*

$$\mathbb{V}\mathrm{ar}\left(\widehat{N}_x(y)\right) = (nhg(x))^{-1} m^{1/2} \kappa(g(x))^2 f_x(y)\psi_1(y) + o((nh)^{-1}m^{1/2}), \tag{4.12}$$

where $\psi_1(y) = [4\pi y(1-y)]^{-1/2}$.

Proof of Proposition 3. From equation (4.7), we know that

$$\widehat{N}_x(y) = \frac{m}{n}\sum_{i=1}^{n} Z_{i,m}.$$

Since $Z_{i,m}$ are $i.i.d$ for each m, thus the variance is given by

$$\mathbb{V}\mathrm{ar}\left(\widehat{N}_x(y)\right) = \frac{m^2}{n}\mathbb{V}\mathrm{ar}(Z_{1,m}),$$

and as we have calculated the $\mathbb{E}\left[\widehat{N}_x(y)\right]$ which is bounded as $n \to \infty, m \to \infty, h \to 0$, we have

$$\mathbb{V}\mathrm{ar}(Z_{1,m}) = \left[\mathbb{E}(Z_{1,m}^2) + O(1)\right].$$

Then,

$$\begin{aligned}\mathbb{E}[Z_{1,m}^2] = \mathbb{E}\Bigg\{ \sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} \Bigg[\mathbb{I}\left(Y_1 \le \frac{k+1}{m}\right) - \mathbb{I}\left(Y_1 \le \frac{k}{m}\right) - m^{-1}f_x(y)\Bigg] \Bigg[\mathbb{I}\left(Y_1 \le \frac{\ell+1}{m}\right) \\ - \mathbb{I}\left(Y_1 \le \frac{\ell}{m}\right) - m^{-1}f_x(y)\Bigg] \times K_h^2(x - X_1)P_{m-1,k}(y)P_{m-1,\ell}(y)\Bigg\}\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\Bigg\{ \sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} \Bigg[\mathbb{I}\left(Y_1 \le \frac{k+1}{m}\right)\mathbb{I}\left(Y_1 \le \frac{\ell+1}{m}\right) + \mathbb{I}\left(Y_1 \le \frac{k}{m}\right)\mathbb{I}\left(Y_1 \le \frac{\ell}{m}\right) \\
&\quad - \mathbb{I}\left(Y_1 \le \frac{k+1}{m}\right)\mathbb{I}\left(Y_1 \le \frac{\ell}{m}\right) - \mathbb{I}\left(Y_1 \le \frac{\ell+1}{m}\right)\mathbb{I}\left(Y_1 \le \frac{k}{m}\right)\Bigg] \\
&\quad \times K_h^2(x - X_1) P_{m-1,k}(y) P_{m-1,\ell}(y) \Bigg\} \\
&\quad - \mathbb{E}\Bigg\{ \sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} \Bigg[m^{-1} f_x(y) \mathbb{I}\left(Y_1 \le \frac{k+1}{m}\right) - m^{-1} f_x(y) \mathbb{I}\left(Y_1 \le \frac{k}{m}\right) - m^{-2} f_x^2(y) \\
&\quad + m^{-1} f_x(y) \mathbb{I}\left(Y_1 \le \frac{\ell+1}{m}\right) - m^{-1} f_x(y) \mathbb{I}\left(Y_1 \le \frac{\ell}{m}\right)\Bigg] \\
&\quad \times K_h^2(x - X_1) P_{m-1,k}(y) P_{m-1,\ell}(y) \Bigg\} \\
&= \mathbb{E}\Bigg\{ \sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} \mathbb{I}\left(Y_1 \le \min\left(\frac{k+1}{m}, \frac{\ell+1}{m}\right)\right) K_h^2(x - X_1) P_{m-1,k}(y) P_{m-1,\ell}(y) \Bigg\} \\
&\quad + \mathbb{E}\Bigg\{ \sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} \mathbb{I}\left(Y_1 \le \min\left(\frac{k}{m}, \frac{\ell}{m}\right)\right) K_h^2(x - X_1) P_{m-1,k}(y) P_{m-1,\ell}(y) \Bigg\} \\
&\quad - \mathbb{E}\Bigg\{ \sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} \mathbb{I}\left(Y_1 \le \min\left(\frac{k+1}{m}, \frac{\ell}{m}\right)\right) K_h^2(x - X_1) P_{m-1,k}(y) P_{m-1,\ell}(y) \Bigg\} \\
&\quad - \mathbb{E}\Bigg\{ \sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} \mathbb{I}\left(Y_1 \le \min\left(\frac{k}{m}, \frac{\ell+1}{m}\right)\right) K_h^2(x - X_1) P_{m-1,k}(y) P_{m-1,\ell}(y) \Bigg\} \\
&\quad - A_{5,m} \\
&= A_{1,m} + A_{2,m} - A_{3,m} - A_{4,m} - A_{5,m}.
\end{aligned}
$$

For $A_{1,m}$, we have

$$
A_{1,m} = \mathbb{E}\Bigg\{ \sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} \mathbb{I}\left(Y_1 \le \min\left(\frac{k+1}{m}, \frac{\ell+1}{m}\right)\right) K_h^2(x - X_1) P_{m-1,k}(y) P_{m-1,l}(y) \Bigg\}
$$

$$
\begin{aligned}
&= \sum_{k=0}^{m-1} \mathbb{E}\left[\mathbb{I}\left(Y_1 \leq \frac{k+1}{m}\right) K_h^2(x - X_1)\right] P_{m-1,k}^2(y) \\
&\quad + 2 \sum_{0 \leq k < \ell \leq m-1} \sum \mathbb{E}\left[\mathbb{I}\left(Y_1 \leq \frac{k+1}{m}\right) K_h^2(x - X_1)\right] P_{m-1,k}(y) P_{\ell-1,k}(y) \\
&= \sum_{k=0}^{m-1} \Gamma_{k+1,m} P_{m-1,k}^2(y) + 2 \sum_{0 \leq k < \ell \leq m-1} \sum \Gamma_{k+1,m} P_{m-1,k}(y) P_{\ell-1,k}(y),
\end{aligned}
$$

where $\Gamma_{k+1,m} = \mathbb{E}[\mathbb{I}(Y_1 \leq \frac{k+1}{m}) K_h^2(x - X_1)]$. Then, $A_{2,m}$, $A_{3,m}$ and $A_{4,m}$ are handled in the same way,

$$
\begin{aligned}
A_{4,m} &= \mathbb{E}\left\{ \sum_{k=0}^{m-1} \sum_{\ell=0}^{m-1} \mathbb{I}\left(Y_1 \leq \min\left(\frac{k}{m}, \frac{\ell+1}{m}\right)\right) K_h^2(x - X_1) P_{m-1,k}(y) P_{m-1,v}(y) \right\} \\
&= \sum_{k=0}^{m-1} \mathbb{E}\left[\mathbb{I}\left(Y_1 \leq \frac{k}{m}\right) K_h^2(x - X_1)\right] P_{m-1,k}^2(y) \\
&\quad + 2 \sum_{0 \leq k < \ell \leq m-1} \sum E\left[\mathbb{I}\left(Y_1 \leq \frac{k}{m}\right) K_h^2(x - X_1)\right] P_{m-1,k}(y) P_{m-1,\ell}(y) \\
&= \sum_{k=0}^{m-1} \Gamma_{k,m} P_{m-1,k}^2(y) + 2 \sum_{0 \leq k < \ell \leq m-1} \sum \Gamma_{k,m} P_{m-1,k}(y) P_{\ell-1,k}(y).
\end{aligned}
$$

From Belalia et al. (2017), we have the result of $A_{2,m}$, which is

$$
\begin{aligned}
A_{2,m} &= \mathbb{E}\left\{ \sum_{k=0}^{m-1} \sum_{\ell=0}^{m-1} \mathbb{I}\left(Y_1 \leq \min\left(\frac{k}{m}, \frac{\ell}{m}\right)\right) K_h^2(x - X_1) P_{m-1,k}(y) P_{m-1,\ell}(y) \right\} \\
&= \sum_{k=0}^{m-1} \mathbb{E}\left[\mathbb{I}\left(Y_1 \leq \frac{k}{m}\right) K_h^2(x - X_1)\right] P_{m-1,k}^2(y) \\
&\quad + 2 \sum_{0 \leq k < \ell \leq m-1} \sum E\left[\mathbb{I}\left(Y_1 \leq \frac{k}{m}\right) K_h^2(x - X_1)\right] P_{m-1,k}(y) P_{m-1,\ell}(y) \\
&= \sum_{k=0}^{m-1} \Gamma_{k,m} P_{m-1,k}^2(y) + 2 \sum_{0 \leq k < \ell \leq m-1} \sum \Gamma_{k,m} P_{m-1,k}(y) P_{\ell-1,k}(y),
\end{aligned}
$$

one can find that $A_{4,m} = A_{2,m}$, thus $A_{2,m} - A_{4,m} = 0$, we turn to calculate $A_{3,m}$ and $A_{5,m}$, we have

$$\begin{aligned}
A_{3,m} &= \mathbb{E}\left\{\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1}\mathbb{I}\left(Y_1 \le \min\left(\frac{k+1}{m}, \frac{\ell}{m}\right)\right) K_h^2(x - X_1)P_{m-1,k}(y)P_{\ell-1,k}(y)\right\} \\
&= \sum_{k=0}^{m-1}\mathbb{E}\left[\mathbb{I}\left(Y_1 \le \frac{k}{m}\right) K_h^2(x - X_1)\right] P_{m-1,k}^2(y) \\
&\quad + 2\sum_{0\le k<\ell\le m-1}\sum\mathbb{E}\left[\mathbb{I}\left(Y_1 \le \frac{k+1}{m}\right) K_h^2(x - X_1)\right] P_{m-1,k}(y)P_{m-1,\ell}(y) \\
&= \sum_{k=0}^{m-1}\Gamma_{k,m}P_{m-1,k}^2(y) + 2\sum_{0\le k<\ell\le m-1}\sum\Gamma_{k+1,m}P_{m-1,k}(y)P_{\ell-1,k}(y),
\end{aligned}$$

$$\begin{aligned}
A_{5,m} &= \mathbb{E}\Bigg\{\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1}\Bigg[m^{-1}f_x(y)\mathbb{I}\left(Y_1 \le \frac{k+1}{m}\right) - m^{-1}f_x(y)\mathbb{I}\left(Y_1 \le \frac{k}{m}\right) - m^{-2}f_x^2(y) \\
&\quad + m^{-1}f_x(y)\mathbb{I}\left(Y_1 \le \frac{\ell+1}{m}\right) - m^{-1}f_x(y)\mathbb{I}\left(Y_1 \le \frac{\ell}{m}\right)\Bigg] \\
&\quad \times K_h^2(x - X_1)P_{m-1,k}(y)P_{m-1,\ell}(y)\Bigg\} \\
&= \mathbb{E}\left\{\sum_{k=0}^{m-1} m^{-1}f_x(y)\left[\mathbb{I}\left(Y_1 \le \frac{k+1}{m}\right) - \mathbb{I}\left(Y_1 \le \frac{k}{m}\right)\right] K_h^2(x - X_1)P_{m-1,k}(y)\right\} \\
&\quad + \mathbb{E}\left\{\sum_{\ell=0}^{m-1} m^{-1}f_x(y)\left[\mathbb{I}\left(Y_1 \le \frac{\ell+1}{m}\right) - \mathbb{I}\left(Y_1 \le \frac{\ell}{m}\right)\right] K_h^2(x - X_1)P_{m-1,\ell}(y)\right\} \\
&\quad - \mathbb{E}\left\{\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} m^{-2}f_x^2(y)K_h^2(x - X_1)P_{m-1,k}(y)P_{m-1,\ell}(y)\right\} \\
&= 2m^{-1}f_x(y)\sum_{k=0}^{m-1}\mathbb{E}\left\{\left[\mathbb{I}\left(Y_1 \le \frac{k+1}{m}\right) - \mathbb{I}\left(Y_1 \le \frac{k}{m}\right)\right] K_h^2(x - X_1)\right\}P_{m-1,k}(y) \\
&\quad - m^{-2}f_x^2(y)\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1}\mathbb{E}\left\{K_h^2(x - X_1)\right\}P_{m-1,k}(y)P_{m-1,\ell}(y)
\end{aligned}$$

$$= 2m^{-1} f_x(y) \sum_{k=0}^{m-1} \left[\Gamma_{k+1,m} - \Gamma_{k,m}\right] P_{m-1,k}(y)$$
$$- m^{-2} f_x^2(y) \sum_{k=0}^{m-1} \sum_{\ell=0}^{m-1} \left\{ h^{-1}\kappa g(x) + O(h) \right\} P_{m-1,k}(y) P_{m-1,\ell}(y).$$

Finally we have,

$$\mathbb{E}\left[Z_{1,m}^2\right] = A_{1,m} - A_{3,m} - A_{5,m}$$
$$= \sum_{k=0}^{m-1} \left[\Gamma_{k+1,m} - \Gamma_{k,m}\right] P_{m-1,k}^2(y) - 2m^{-1} f_x(y) \sum_{k=0}^{m-1} \left[\Gamma_{k+1,m} - \Gamma_{k,m}\right] P_{m-1,k}(y)$$
$$- m^{-2} f_x^2(y) \sum_{k=0}^{m-1} \sum_{\ell=0}^{m-1} \left\{ h^{-1}\kappa g(x) + O(h) \right\} P_{m-1,k}(y) P_{m-1,\ell}(y).$$

Similar calculation as Equation (21) in Belalia et al. (2017), we have

$$\Gamma_{k+1,m} = \mathbb{E}\left[K_h^2(x - X_1) \mathbb{E}\left\{ \mathbb{I}\left(Y_1 \leq \frac{k+1}{m} \right) \middle| X_1 \right\} \right]$$
$$= \mathbb{E}\left[K_h^2(x - X_1) F_{X_1}\left(\frac{k+1}{m} \right) \right]$$
$$= h^{-2} \int g(z) F_z\left(\frac{k+1}{m} \right) K^2\left(\frac{z-x}{h} \right) dz$$
$$= h^{-1} \int g(x + hv) F_{x+hv}\left(\frac{k+1}{m} \right) K^2(v) dv$$
$$= h^{-1} \int \left[g(x) + hvg'(x) + o(h) \right] \left[F_x\left(\frac{k+1}{m} \right) + hv F_x^{(1,0)}\left(\frac{k+1}{m} \right) + o(h) \right] K^2(v) dv$$
$$= h^{-1}\kappa g(x) F_x\left(\frac{k+1}{m} \right) + O\left(h\gamma\left(\frac{k+1}{m} \right) \right).$$

One can use Taylor expansion to get

$$F_x\left(\frac{k+1}{m}\right) = F_x\left(\frac{k}{m}\right) + m^{-1} f_x\left(\frac{k}{m}\right) + o(m^{-1}),$$

and expand $\gamma\left(\frac{k+1}{m}\right)$ too, then we have

$$\begin{aligned}\sum_{k=0}^{m-1}\left[\Gamma_{k+1,m} - \Gamma_{k,m}\right] P_{m-1,k}^2(y) = &\sum_{k=0}^{m-1} h^{-1}\kappa g(x)\left[m^{-1} f_x\left(\frac{k}{m}\right) + o(m^{-1})\right] P_{m-1,k}^2(y) \\ &+ \sum_{k=0}^{m-1} o(m^{-1}h) P_{m-1,k}^2(y). \end{aligned} \tag{4.13}$$

According to Lemma 3, we have

$$\sum_{k=0}^{m-1} F_x\left(\frac{k}{m}\right) P_{m-1,k}^2(y) = F_x(y) S_{m-1}(y) + O(I_{m-1}(y)),$$

where $I_{m-1}(y) = \sum_{k=0}^{m-1} |k/m - y| P_{m-1,k}^2(y) = O_y(m^{-3/4})$. Then

$$\sum_{k=0}^{m-1} m^{-1} f_x\left(\frac{k}{m}\right) P_{m-1,k}^2(y) = m^{-1}\{f_x(y) S_{m-1}(y) + O(I_{m-1}(y))\},$$

substituting the result into equation (4.13), and using Lemma 3, we obtain

$$\begin{aligned}\sum_{k=0}^{m-1}\left[\Gamma_{k+1,m} - \Gamma_{k,m}\right] P_{m-1,k}^2(y) &= h^{-1}\kappa g(x)[m^{-1} f_x(y) S_{m-1}(y) + O(m^{-1} I_{m-1}(y)) \\ &\quad + o(m^{-1} S_{m-1}(y)) + o(m^{-1} I_{m-1}(y))] + o(m^{-1} h S_{m-1}(y)) \\ &= h^{-1}\kappa g(x)\Big\{m^{-1} f_x(y) m^{-1/2}[\psi_1(y) + o(1)] + O(m^{-1} m^{-3/4}) \\ &\quad + o\left(m^{-1} m^{-1/2}[\psi_1(y) + o(1)]\right)\Big\} + o\left(m^{-1} h m^{-1/2}[\psi_1(y) + o(1)]\right) \\ &= h^{-1}\kappa g(x)\left[m^{-1} f_x(y) m^{-1/2}\psi_1(y)\right] + o(h^{-1} m^{-3/2}) + O(h^{-1} m^{-7/4})\end{aligned}$$

$$
\begin{aligned}
&+ o(h^{-1}m^{-3/2}) + o(hm^{-3/2}) \\
&= h^{-1}\kappa g(x)\left[m^{-1}f_x(y)m^{-1/2}\psi_1(y)\right] + o(h^{-1}m^{-3/2}),
\end{aligned} \tag{4.14}
$$

where $\psi_1(y) = [4\pi y(1-y)]^{-1/2}$. Similarly, we have

$$
\begin{aligned}
\sum_{k=0}^{m-1}\left[\Gamma_{k+1,m} - \Gamma_{k,m}\right] P_{m-1,k}(y) &= \sum_{k=0}^{m-1} h^{-1}\kappa g(x)\left[m^{-1}f_x\left(\frac{k}{m}\right) + o(m^{-1})\right] P_{m-1,k}(y) + o(m^{-1}h) \\
&= h^{-1}m^{-1}\kappa g(x) f_x(y) + o(h^{-1}m^{-1}).
\end{aligned}
$$

As for the variance of $\widehat{N}_x(y)$, we have

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\widehat{N}_x(y)\right] &= \frac{m^2}{n}\mathbb{E}[Z_{1,m}^2] \\
&= n^{-1}m^2\left[h^{-1}\kappa g(x)\left[m^{-1}f_x(y)m^{-1/2}\psi_1(y)\right] + o(h^{-1}m^{-3/2})\right. \\
&\quad - 2m^{-1}f_x(y)\left[h^{-1}m^{-1}\kappa g(x) f_x(y) + o(h^{-1}m^{-1})\right] \\
&\quad \left. - m^{-2}f_x^2(y)\left(h^{-1}\kappa g(x) + O(h)\right)m^{-1/2}[\psi_1(y) + o(1)]\right] \\
&= n^{-1}m^2\left[h^{-1}m^{-3/2}\kappa g(x) f_x(y)\psi_1(y) + o(h^{-1}m^{-3/2})\right. \\
&\quad - 2h^{-1}m^{-2}\kappa g(x) f_x^2(y) - o(h^{-1}m^{-2}) - h^{-1}m^{-5/2}\kappa g(x) f_x^2(y)\psi_1(y) \\
&\quad \left. - O(hm^{-5/2}) - o(h^{-1}m^{-5/2})\right] \\
&= (nh)^{-1}m^{1/2}\kappa g(x) f_x(y)\psi_1(y) + o((nh)^{-1}m^{1/2}),
\end{aligned}
$$

where ψ_1(y) is defined as previous. □

Theorem 4.2. *Under Assumptions 1-3, and assuming* $(nh)^{-1}m \to 0$*, we have for*

$y \in (0,1)$ *that*

$$\mathbb{AV}ar[\widehat{f}_{x,mh}(y)] = (nhg(x))^{-1} m^{1/2} \kappa f_x(y) \psi_1(y), \tag{4.15}$$

where $\psi_1(y) = [4\pi y(1-y)]^{-1/2}$.

Now, we can calculate the asymptotic integrated mean squared error (AIMSE) for each fixed x by using following equation

$$\text{IMSE}(\widehat{f}_{x,mh}(y)) = \int_0^1 \mathbb{E}\left(\widehat{f}_{x,mh}(y) - f_x(y)\right)^2 dy.$$

Corollary 1. *Under the assumption of Theorem 4.1 and Theorem 4.2, we have*

$$\begin{aligned}\text{AIMSE}\left(\widehat{f}_{x,mh}(y)\right) = m^{-2}\left(C_2 + 4^{-1}C_0 - C_1\right) + m^{-1}h^2\kappa_2\Bigg(\left[\frac{g'(x)}{g(x)}\right]F_1 - 2\left[\frac{g'(x)}{g(x)}\right]E_1 \\ - E_2 + 2^{-1}F_2\Bigg) + h^4\kappa_2^2\left(4^{-1}D_2 + \left[\frac{g'(x)}{g(x)}\right]G + \left[\frac{g'(x)}{g(x)}\right]^2 D_1\right) \\ + (nhg(x))^{-1}m^{1/2}\kappa H,\end{aligned} \tag{4.16}$$

for $i = 0, 1, 2$ *and* $j = 1, 2$, *where*

$$\begin{aligned} C_i &= \int_0^1 y^i \left[f_x'(y)\right]^2 dy, \qquad D_j = \int_0^1 \left[F_x^{(j,1)}(y)\right]^2 dy, \\ E_j &= \int_0^1 y f_x'(y) F_x^{(j,1)}(y) dy, \quad F_j = \int_0^1 f_x'(y) F_x^{(j,1)}(y) dy, \\ G &= \int_0^1 F_x^{(1,1)}(y) F_x^{(2,1)}(y) dy, \quad H = \int_0^1 f_x(y)\psi_1(y) dy, \end{aligned}$$

and where $\psi_1(y)$ *is defined as in Theorem 4.2.*

We point out that with similarity to the estimators based on kernel method, this AIMSE can be minimized with respect to (m, h) to select the optimal choice of the

bandwidth parameters m, h.

4.4 Asymptotic Normality

In order to establish the asymptotic normality for the proposed two-stage estimators. We first, derive the distribution limit of $\widehat{N}_x(y)$, and consequently obtain that of $\widehat{f}_{x,mh}$.

Proposition 4. *Under the Assumption* $1-3$*, assuming* $(nh)^{-1}m \to 0$*, we have for* $y \in (0,1)$ *that*

$$(nh)^{1/2}m^{-1/4}\Big\{\widehat{N}_x(y) - g(x)\Big[-m^{-1}yf'_x(y) + (2m)^{-1}f'_x(y) + h^2\kappa_2\frac{g'(x)}{g(x)}F_x^{(1,1)}(y) + \frac{h^2\kappa_2}{2}F_x^{(2,1)}(y)\Big]\Big\}$$
$$\xrightarrow{\mathcal{D}} N\left(0, \kappa\frac{(g(x))^2}{g(x)}\left[f_x(y)\psi_1(y)\right]\right), \tag{4.17}$$

where $\psi_1(y)$ *defined as in Proposition 3 and "*$\xrightarrow{\mathcal{D}}$*" denotes convergence in distribution.*

Proof of Proposition 4. We know that

$$\widehat{N}_x(y) = \frac{m}{n}\sum_{i=0}^{n} Z_{i,m}$$

under the condition that the random variables $Z_{1,m}, ..., Z_{n,m}$ are *i.i.d.*, thus $Z_{i,m}$ is an average of the *i.i.d.* random variables. Then we can use the central limit theorem for double arrays (e.g. (Serfling, 2002, Section 1.9.3)), that means if the following Lindberg condition holds we can have the desired asymptotic normality of $\widehat{N}_x(y)$,

$$A_n = \frac{1}{s_m^2}\mathbb{E}\left\{[Z_{1,m} - \mathbb{E}(Z_{1,m})]^2\mathbb{I}\left(|Z_{1,m} - \mathbb{E}(Z_{1,m})| > \epsilon s_m n^{1/2}\right)\right\} \to 0, \tag{4.18}$$

for every $\epsilon > 0$, as $n \to \infty$, where $s_m^2 = \mathbb{V}\text{ar}(Z_{1,m})$ is given by (4.14). Following the

idea of Babu et al. (2002, Proof of Proposition 1), we observe that

$$\begin{aligned}
|Z_{1,m}| &= \left| \sum_{k=0}^{m-1} \left[\mathbb{I}\left(Y_1 \leq \frac{k+1}{m}\right) - \mathbb{I}\left(Y_1 \leq \frac{k}{m}\right) - m^{-1} f_x(y) \right] P_{m-1,k}(y) K_h(x - X_1) \right| \\
&\leq \max_{0 \leq k \leq m-1} \left(P_{m-1,k}(y) K_h(x - X_1) + m^{-1} f_x(y) K_h(x - X_1) \right) \\
&\leq \left(\sum_{k=0}^{m-1} P_{m-1,k}^2(y) \right)^{1/2} h^{-1} M_K + m^{-1} M_f h^{-1} M_K \\
&= O(h^{-1} m^{-1/4}) + O(h^{-1} m^{-1}) \\
&= O\left(h^{-1} m^{-1/4}\right), \qquad (4.19)
\end{aligned}$$

where M_K, M_f is such that $K(x) \leq M_K$ and $f_x(y) \leq M_f$ respectively. From the equation (4.6), we have

$$\mathbb{E}(Z_{1,m}) = m^{-1} \mathbb{E}(\widehat{N}_x(y)) = O(m^{-2}) + O(h^2 m^{-1}),$$

then

$$|Z_{1,m} - \mathbb{E}(Z_{1,m})| \leq O\left(h^{-1} m^{-1/4}\right) + O(m^{-2}) + O(h^2 m^{-1}) = O\left(h^{-1} m^{-1/4}\right).$$

Then for checking the Lindberg condition, we have

$$\begin{aligned}
\frac{|Z_{1,m} - \mathbb{E}(Z_{1,m})|}{s_m n^{1/2}} &\leq \frac{O\left(h^{-1} m^{-1/4}\right)}{s_m n^{1/2}} \\
&= \frac{O\left(h^{-1} m^{-1/4}\right)}{n^{1/2} (h g(x))^{-1/2} m^{-3/4} (\kappa f_x(y) \psi_1(y))^{1/2}} = O\left(\left((nh)^{-1} m\right)^{1/2}\right),
\end{aligned}$$

and we notice that when $m \to \infty$, $n \to \infty$, $nh \to \infty$ and $(nh)^{-1} m \to 0$, then

$A_n \to 0$, which completes the proof. □

Theorem 4.3. *Under Assumption 1-3, assuming* $(nh)^{-1}m \to 0$, *we have for* $y \in (0,1)$ *that*

$$(nh)^{1/2}m^{-1/4}\Big\{\widehat{f}_{x,mh}(y) - f_x(y) - \Big[-m^{-1}yf_x'(y) + \frac{1}{2}m^{-1}f_x'(y) + h^2\kappa_2\frac{g'(x)}{g(x)}F_x^{(1,1)}(y) + \frac{h^2\kappa_2}{2}F_x^{(2,1)}(y)\Big]\Big\} \xrightarrow{\mathcal{D}} N\left(0, \kappa\frac{f_x(y)\psi_1(y)}{g(x)}\right),$$

where $\psi_1(y) = [4\pi y(1-y)]^{-1/2}$.

Note that, additionally, under the condition $nh^5 \to 0$, we have

$$(nh)^{1/2}m^{-1/4}\left(\widehat{f}_{x,mh}(y) - f_x(y)\right) \xrightarrow{\mathcal{D}} N\left(0, \kappa\frac{f_x(y)\psi_1(y)}{g(x)}\right). \tag{4.20}$$

We can construct a $100\%(1-\alpha)$ confidence interval as the follows

$$\left[\widehat{f}_{x,mh}(y) - z_{1-\alpha/2}\sqrt{\frac{m^{1/2}\kappa f_x(y)\psi_1(y)}{nhg(x)}},\quad \widehat{f}_{x,mh}(y) + z_{1-\alpha/2}\sqrt{\frac{m^{1/2}\kappa f_x(y)\psi_1(y)}{nhg(x)}}\right].$$

4.5 Simulation study

We observe $(X_1, Y_1), ..., (X_n, Y_n)$ that are independently identically distributed random vectors. The variables X_i are assumed to be distributed uniformly on $[0,1]$ and Y_i conditioned on $X_i = x$ has the density

$$f_x(y) = \frac{y^{10x}(1-y)^4}{B(10x+1,5)} \tag{4.21}$$

for $0 < y < 1$, where $B(\cdot,\cdot)$ stands for the beta function. Note that, the conditional

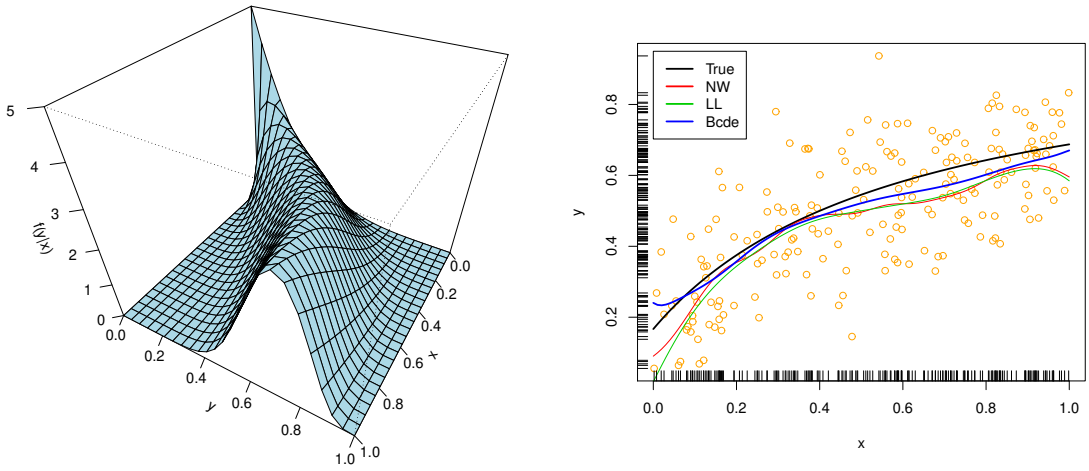
mean of this distribution is given

$$r(x) = \mathbb{E}\left[Y \mid X = x\right] = \int_0^1 y\, f_x(y)\, \mathrm{d}y = \frac{10x+1}{10x+6}. \tag{4.22}$$

The latter can be estimated using the plug-in approach and the conditional density estimator (4.3), namely, simple algebra leads to

$$\hat{r}_{m,n}(x) = \int_0^1 y\, \widehat{f}_{x,mh}(y)\, \mathrm{d}y = \sum_{k=0}^{m-1} \frac{k+2}{m+1} \left[\hat{F}_{x,h}\Big([k+1]/m\Big) - \hat{F}_{x,h}(k/m)\right]. \tag{4.23}$$

The shape of this conditional density function is shown in Figure 4.1a (True). A Typical sample of size $n = 200$ from Model (4.21) with the true curve of the regression function (4.22)(black line), The Bernstein estimator (4.23)(blue line and $m = 25$), the Nadaraya-Watson estimator(Red line), and local linear estimator(green line) are depicted in Figure 4.1b, as one can see the Bernstein regression estimator is more closer than to the true regression curve that the NW, and LL estimator, in particular at the boundaries of the support of X.



(a) (b)

Figure 4.1: **Left:** The true conditional density of model (4.21). **Right:** Typical sample of size $n = 200$ from Model (4.21) with the true curve of the regression function (4.22)(black line), The Bernstein estimator (4.23)(blue line and $m = 25$), the Nadaraya-Watson estimator(Red line), and local linear estimator(green line).

To study the finite-sample behaviour of the proposed estimator (Bcde) (4.3) compared to that of (**NW**) and (**LL**), a $B = 500$ samples of sizes $n = 50, 100, 150, 200, 250, 500$ were generated form Model (4.21). On each sample the estimators **Bcde**, **NW** and **LL** were calculated. Further, we evaluated the global properties of these estimators in terms of the integrated mean square error (IMSE)

$$I(f) = \int_x \int_y \mathbb{E}\left[\hat{f}_x(y) - f_x(y)\right]^2 dydx \tag{4.24}$$

where the integrals are approximated by a 50×50 grid on (y, x) and $\hat{f}_x(y)$ representing an estimate of the true conditional density function $f_x(y)$. The estimators, **NW** and **LL**, and their bandwidth parameters are obtained using the function `cde` in the **R** package `hdrcde`. This function selects the best bandwidth parameter in terms of

IMSE. However, for a fair comparison, the best bandwidth in direction of X was calculated automatically using that function and a grid of value of $h_y = m/350$. For the **Bcde**, a grid of values of m, from $m = 5$ to $m = 80$ spaced by 5 was taken. Figure 4.2 illustrates the IMSE of Bernstein estimator as a function of m and provides the IMSE for the two competitors. First, we see that the IMSE decreases, and the optimal bandwidth parameter increases as the sample size n increases. Second, for all sample size, our estimator outperforms Nadaraya-Watson estimator. Third, for small and moderate sample size ($n = 50, 100$), the optimal IMSE of Bernstein estimator is better than that of the local linear and their performance in terms of IMSE is comparable for large sample size. We point out that in Figure 4.2, another version of the proposed estimator based on the local linear conditional distribution function estimate(dark green dotted line) smoothed in the first stage was also added.

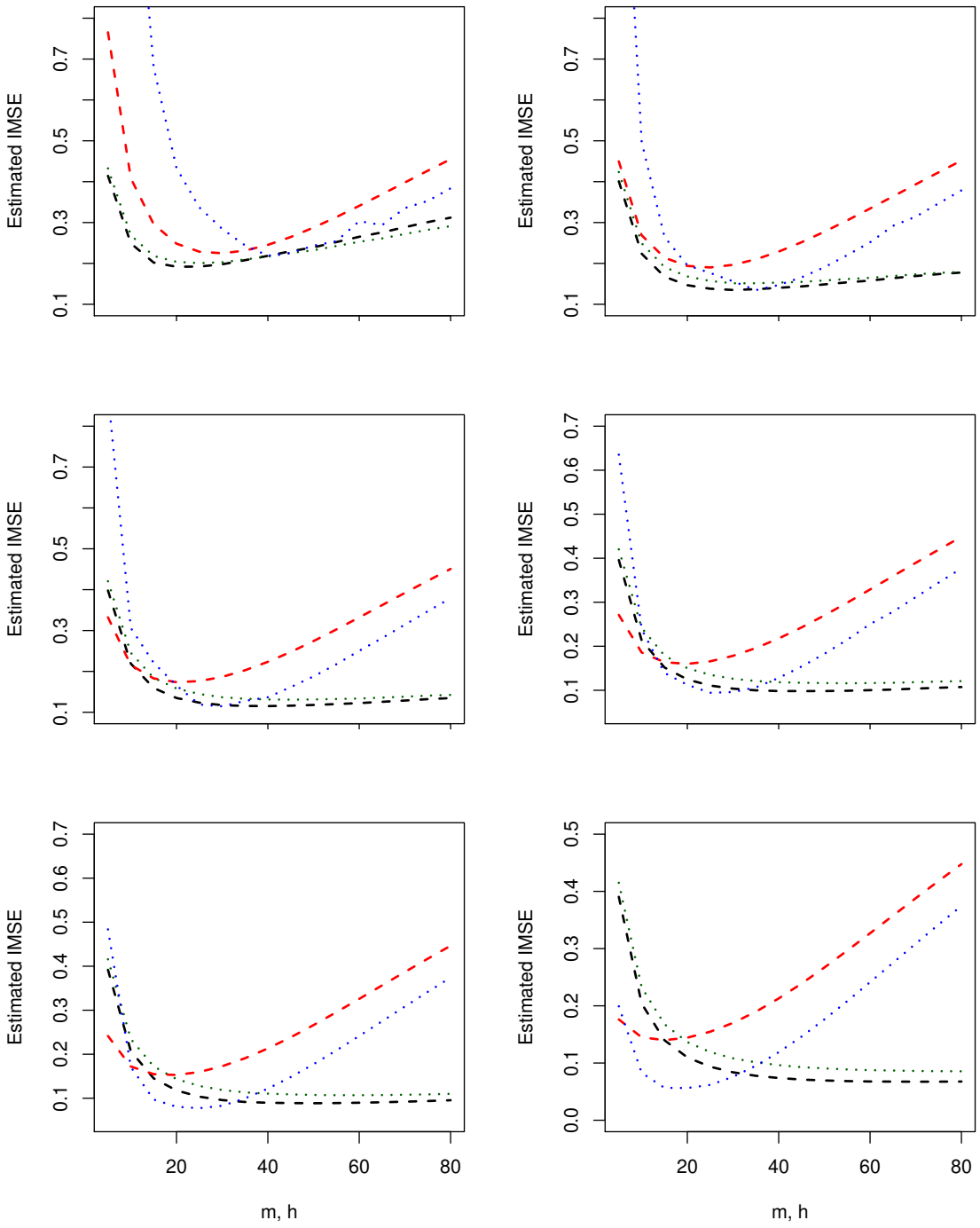


Figure 4.2: The estimate integrated mean square error as a function of $m, h = m/350$ for Bernstein estimator Bcde (black and dark green lines) plotted with the local polynomial estimators (red dashed red line corresponds to NW and blue dotted line to LL). The sample size was taken to be $n = 50, 100$ (first row), $n = 150, 200$ (second row), $n = 250, 500$ (third row).

4.6 Old Faithful Data Application

We apply the Bernstein conditional density estimator on the Old Faithful Geyser data, which is the data of waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. This data set is firstly analyzed by Azzalini and Bowman (1990) and then is widely used in the nonparametric statistics for real data application, for example, see the work of Silverman (1986) for comparing density estimates, Di Lucca et al. (2013) for Bayesian nonparametric auto-regression model and Matzner-Løber et al. (1998) for nonparametric forecasting. The data has 272 observations and 2 variables depicted in Figure 4.3 with the estimated regression function. Also, we plot the estimators (4.1) and (4.3) for eruption duration conditional on waiting time. We can notice that the they capture the information of the data very well, especially, the bi-modality of the data captured with the Bernstein conditional density estimator.

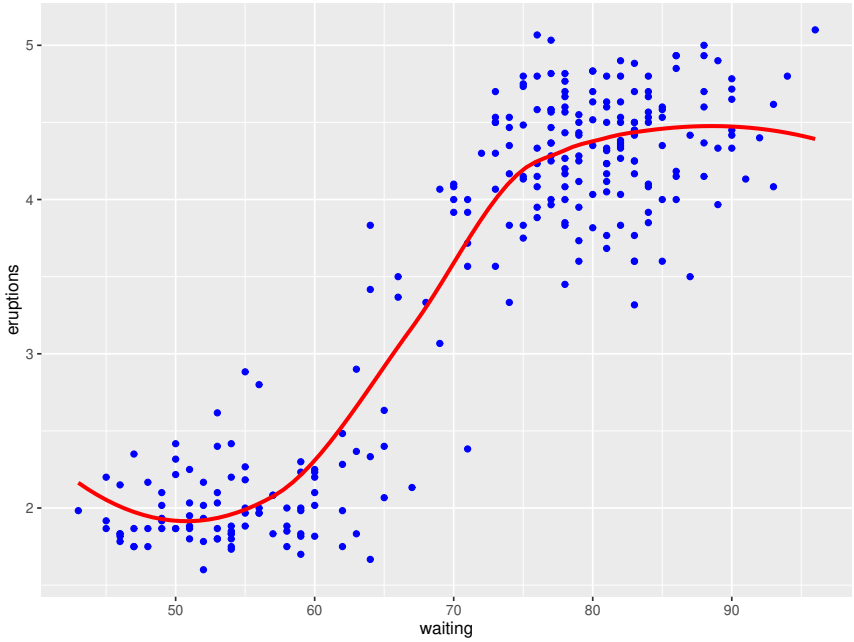
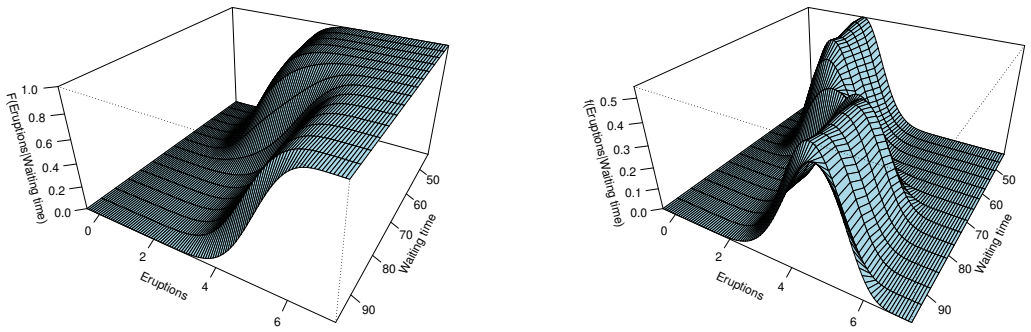


Figure 4.3: Eruptions duration against waiting time with estimated regression curve using the Bernstein estimator (4.23).



(a) (b)

Figure 4.4: Bernstein estimates of the distribution of eruption duration conditional on waiting time; (a) the conditional density ($m = 25$), (b) the conditional distribution function ($m = 25$).

Chapter 5

Conclusions and Further Questions

In this thesis, we have discussed nonparametric estimation through kernel or Bernstein polynomials based methods with focusing on the conditional density estimate. Simulation study have shown some performance of the Bernstein-type estimators compared to the kernel-type estimators for an appropriate choice of the polynomials order m. Besides, it is well-known that the bandwidth parameter h has dominating influence on the behaviour of kernel-type estimators. Many techniques have been derived for this process, such as cross-validation, plug-in and normal reference method. Also, it the case of the proposed estimator, which can be affected by the choice of two bandwidth parameters (h, m). A suggested selection method can be done by minimizing the integrated mean square error with respect to (h, m) using cross-validation approach.

Further, an extension of the proposed estimator to multivariate predictor case, which will make it more flexible and adaptive for practical implementation is left for future work . Moreover, the work of Xian (2005) pointed out that, Bernstein polynomials

can be transformed as in the examples of Feller (1971, Lemma 1, Section VII.1). For instance, we can have a polynomial with Poisson distribution based on Bernstein polynomials. Indeed, using the notation in Theorem 3.2, for a Poisson distribution with parameter λ, let $\lambda = mx$ and $m \to \infty$, we have

$$P_m(f)(x) = \exp(-mx) \sum_{k=0}^{\infty} f\left(\frac{k}{m}\right) \frac{(mx)^k}{k!} \to f(x)$$

uniformly in every finite x-interval. Using this type of polynomials to play the smoothing role in the first stage can be an alternative approach.

Appendices

Appendix A

Supplementary materials

A.1 Indicator Function

Definition A.1. *Let Ω be a sample space and $E \subseteq \Omega$ be an event. The indicator function of the event E is a random variable defined as follows:*

$$\mathbb{I}_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}$$

where ω indicate a event, for simplicity we denote $\mathbb{I}_E(\omega)$ by $\mathbb{I}(E)$.

The indicator function is widely used in nonparametric statistics with following basic properties.

- The n_{th} power of $\mathbb{I}(E)$ is equal to $\mathbb{I}(E)$,

$$(\mathbb{I}(E))^n = \mathbb{I}(E).$$

- The expected value of $\mathbb{I}_E$ is equal to $\mathbb{P}(E)$,

$$\begin{aligned}\mathbb{E}\left(\mathbb{I}(E)\right) &= \sum_x x\mathbb{P}(x) \\ &= 1 \cdot \mathbb{P}(1) + 0 \cdot \mathbb{P}(0) \\ &= 1 \cdot \mathbb{P}(E) + 0 \cdot \mathbb{P}(E^c) \\ &= \mathbb{P}(E).\end{aligned}$$

- The variance of $\mathbb{I}(E)$ is equal to $\mathbb{P}(E)(1 - \mathbb{P}(E))$,

$$\begin{aligned}\mathbb{V}\text{ar}\left(\mathbb{I}(E)\right) &= \mathbb{E}\left(\left(\mathbb{I}(E)\right)^2\right) - \mathbb{E}\left(\mathbb{I}(E)\right)^2 \\ &= \mathbb{E}\left(\mathbb{I}(E)\right) - \mathbb{E}\left(\mathbb{I}(E)\right)^2 \\ &= \mathbb{P}(E) - \mathbb{P}(E)^2 \\ &= \mathbb{P}(E)(1 - \mathbb{P}(E)).\end{aligned}$$

- Intersections. If E and F are two events, then

$$\mathbb{I}\left(E \bigcap F\right) = \mathbb{I}(E)\mathbb{I}(F).$$

Because if $E \bigcap F$ happens then $\mathbb{I}\left(E \bigcap F\right) = 1$, we have E and F both happen, then $\mathbb{I}(E)\mathbb{I}(F) = 1$; if $E \bigcap F$ does not happen, then $\mathbb{I}\left(E \bigcap F\right) = 0$, that means E or F does not happen, then $\mathbb{I}(E)\mathbb{I}(F) = 0$.

A.2 Empirical Distribution Function Properties

At any fixed value x, we have

$$\begin{aligned}\mathbb{E}\left(F_n(x)\right) &= \frac{1}{n}\sum_{1}^{n}\mathbb{E}\left(\mathbb{I}(X_i \le x)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(X < x)\\ &= F(x).\end{aligned}$$

which means $F_n(x)$ is an unbiased estimator for $F(x)$. And

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(F_n(x)\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}\left(\mathbb{I}(X_i \le x)\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left(F(x)(1-F(x))\right)\\ &= \frac{1}{n}F(x)(1-F(x)).\end{aligned}$$

The convergence property of empirical distribution function is based on following theorems.

Theorem A.1. *(Strong Law of Large Number) Let $(X_n)_{n\ge1}$ be a sequence of independent and identically distributed (i.i.d.) random variables with $\mathbb{E}(X_1^4) < \infty$ and $\mathbb{E}(X_1) = \mu$. Then*

$$\frac{S_n}{n} := \frac{1}{n}\sum_{i=1}^{n}X_i \to \mu \qquad \textit{almost surely.}$$

For a given $x \in \mathbb{R}$, we can apply the strong law of large number to the sequence

$\mathbb{I}(X_i < x), i = 1, \ldots, n$ to assert that

$$F_n(x) \to F(x)$$

almost surely, because $\mathbb{E}[|\mathbb{I}(X_i < x)|] < \infty$.

In this case, $F_n(x)$ is a reasonable estimate of $F(x)$ for a given $x \in \mathbb{R}$. But when $F_n(x)$ and $F(x)$ both are viewed as function of x, the strong law of large number cannot be applied.

Theorem A.2. *(Glivenko-Cantelli) Let $X_1, X_2, \ldots X_n$ be a collection of i.i.d. random variables with cdf F, and let $F_n(x)$ denote the empirical distribution function. Then as $n \to \infty$,*

$$\mathbb{P}\left[\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0\right] = 1,$$

or equivalently

$$\mathbb{P}\left[\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0\right] = 1,$$

that is, the convergence is uniform in x.

Proof. Let $\epsilon > 0$, then fix $k > 1/\epsilon$, and consider "knot" points $\kappa_0, \ldots, \kappa_k$ such that

$$-\infty = \kappa_0 < \kappa_1 \leq \kappa_2 \leq \cdots \leq \kappa_{k-1} < \kappa_k = \infty,$$

that define a partition of $\mathbb{R}$ into k disjoint intervals such that

$$F(\kappa_j^-) \leq \frac{j}{k} \leq F(\kappa_j) \qquad j = 1, \ldots, k-1.$$

where, for each j,

$$F(k_j^-) = \mathbb{P}[X_j < \kappa_j] = F(\kappa_j) - \mathbb{P}[X = \kappa_j].$$

Then, by construction, if $\kappa_{j-1} < \kappa_j$,

$$F(\kappa_j^-) - F(\kappa_{j-1}) \leq \frac{j}{k} - \frac{(j-1)}{k} = \frac{1}{k} < \epsilon.$$

Recall from the strong law of large number we can write

$$|F_n(\kappa_j) - F(\kappa_j)| \xrightarrow{a.s.} 0 \quad \text{and} \quad |F_n(\kappa_j^-) - F(\kappa_j^-)| \xrightarrow{a.s.} 0$$

as $n \to \infty$. So looking at the maximum over all j, we have

$$\triangle_n = \max_{j=1,\ldots,k-1} \left\{ |F_n(\kappa_j) - F(\kappa_j), F_n(\kappa_j^-) - F(\kappa_j^-)| \right\} \xrightarrow{a.s.} 0$$

as $n \to \infty$.

For any x, find the interval within which x lies, that is, identity j such that

$$\kappa_{j-1} \leq x < \kappa_j.$$

Then we have following inequality hold

$$F_n(x) - F(x) \leq F_n(\kappa_j^-) - F(\kappa_{j-1}) \leq F_n(\kappa_j^-) - F(\kappa_j^-) + \epsilon,$$

and

$$F_n(x) - F(x) \geq F(\kappa_{j-1}) - F_n(\kappa_j^-) \geq F_n(\kappa_{j-1}) - F(\kappa_{j-1}) - \epsilon.$$

Thus for any x,

$$F_n(\kappa_{j-1}) - F(\kappa_{j-1}) - \epsilon \leq F_n(x) - F(x) \leq F_n(\kappa_j^-) - F(\kappa_j^-) + \epsilon,$$

and then

$$|F_n(x) - F(x)| \leq \triangle_n + \epsilon \xrightarrow{a.s.} \epsilon$$

as $n \to \infty$. And

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} \epsilon$$

as $n \to \infty$, which follows

$$\mathbb{P}\left[\lim_{n \to \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0\right] = 1,$$

that completes the proof. □

This result is a very important result in empirical process theory and modern econometrics.

A.3 Naive Density Estimator Properties

We review the statistical properties for the naive density estimator with the method

in Rosenblatt (1956), we have

$$\begin{aligned}\mathbb{E}(F_n(x)F_n(x')) &= \mathbb{E}\left(\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{I}(X_i \le x)\mathbb{I}(X_i \le x')\right)\\ &= \mathbb{E}\left(\frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{I}(X_i \le x)\mathbb{I}(X_i \le x') + \sum\sum_{i\neq j}\mathbb{I}(X_i \le x)\mathbb{I}(X_j \le x')\right)\right)\\ &= \frac{1}{n}F(\min(x,x')) + \frac{n-1}{n}F(x)F(x').\end{aligned}$$

Then

$$\mathbb{C}\mathrm{ov}\left(F_n(x), F_n(x')\right) = \frac{1}{n}\left[F(\min(x,x')) - F(x)F(x')\right],$$

by (2.6), we have

$$\begin{aligned}\mathbb{C}\mathrm{ov}(\hat{f}_{nh}(x), \hat{f}_{nh}(x')) &= \frac{1}{4h^2}\mathbb{C}\mathrm{ov}\left(F_n(x+h) - F_n(x-h), F_n(x'+h) - F_n(x'-h)\right)\\ &= \frac{1}{4nh^2}\Big[F(\min(x+h, x'+h)) - F(x+h)F(x'+h)\\ &\quad - F(\min(x+h, x'-h)) + F(x+h)F(x'-h) - F(\min(x-h, x'+h))\\ &\quad + F(x-h)F(x'+h) + F(\min(x-h, x'-h)) - F(x-h)F(x'-h)\Big].\end{aligned}$$

Set $x = x'$,

$$\mathbb{V}\mathrm{ar}(\hat{f}_{nh}(x)) = \frac{1}{4nh^2}\Big[F(x+h) - F(x-h) - \left(F(x+h) - F(x-h)\right)^2\Big]. \qquad \text{(A.1)}$$

Now we consider the behaviour of $\hat{f}_n(x)$ by evaluating the mean square error (MSE), where x is fixed as $n \to \infty$ and $h \to 0$,

$$\mathbb{E}\left[\hat{f}_{nh}(x) - f(x)\right]^2 = \mathbb{V}\mathrm{ar}\left(\hat{f}_{nh}(x)\right) + \mathbb{B}\mathrm{ias}\left(\hat{f}_{nh}(x)\right)^2$$

$$= \frac{1}{4nh^2}\Big[F(x+h) - F(x-h) - (F(x+h) - F(x-h))^2\Big] + \left[\frac{1}{2h}(F(x+h) - F(x-h)) - f(x)\right]^2. \quad \text{(A.2)}$$

Assuming $F(\cdot)$ is third differentiable, we use the Taylor expansion for the MSE,

$$\begin{aligned} F(x+h) - F(x-h) &= \int_{x-h}^{x+h} f(t)dt \\ &= \int_{x-h}^{x+h} \left[f(x) + f'(x)(t-x) + \frac{f''(x)}{2}(t-x)^2 + O((t-x)^3)\right] dt \\ &= 2hf(x) + \frac{1}{3}h^3 f''(x) + O(h^4), \end{aligned}$$

then

$$\mathbb{E}\left[\hat{f}_{nh}(x) - f(x)\right]^2 \sim \frac{f(x)}{2nh} + \frac{h^4}{36}\left(f''(x)\right)^2 + o\left(\frac{1}{nh} + h^4\right)$$

as $h \to 0$ and $n \to \infty$.

A.4 Kernel Density Estimator Properties

Proof of Theorem 2.2. The mean squared error formula is given as

$$\text{MSE}\left(\hat{f}_{nh}(x)\right) = \mathbb{V}\text{ar}\left(\hat{f}_{nh}(x)\right) + \left[\mathbb{B}\text{ias}\left(\hat{f}_{nh}(x)\right)\right]^2.$$

We will evaluate the $\mathbb{B}\text{ias}\left(\hat{f}_{nh}(x)\right)$ and $\mathbb{V}\text{ar}\left(\hat{f}_{nh}(x)\right)$ terms separately. For the bias calculation we use the Taylor expansion.

The bias term is given by

$$\begin{aligned}
\mathbb{B}\text{ias}\left(\hat{f}_{nh}(x)\right) &= \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right] - f(x) \\
&= h^{-1}\mathbb{E}\left[K\left(\frac{X_1 - x}{h}\right)\right] - f(x) \\
&= h^{-1}\int_{-\infty}^{\infty} f(X_1) K\left(\frac{X_1 - x}{h}\right) dX_1 - f(x) \\
&= h^{-1}\int_{-\infty}^{\infty} f(x + hv) K(v) h dv - f(x) \\
&= \int_{-\infty}^{\infty} \left\{f(x) + f^{(1)}(x)hv + \frac{1}{2}f^{(2)}(x)h^2v^2 + O(h^3)\right\} K(v) dv - f(x) \\
&= \left\{f(x) + 0 + \frac{h^2}{2}f^{(2)}(x)\int_{-\infty}^{\infty} v^2 K(v) dv + O(h^3)\right\} - f(x) \\
&= \frac{h^2}{2}f^{(2)}(x)\int_{-\infty}^{\infty} v^2 K(v) dv + O(h^3), \qquad \text{(A.3)}
\end{aligned}$$

where the $O(h^3)$ term comes from

$$(1/3!)h^3\left|\int_{-\infty}^{\infty} f^{(3)}(\tilde{x}) v^3 K(v)\right| dv \leq Ch^3 \int_{-\infty}^{\infty} \left|v^3 K(v) dv\right| = O(h^3),$$

where C is a positive constant, and where $\tilde{x}$ lies between x and $x + hv$.

Next we consider the variance term, observe that

$$\begin{aligned}
\mathbb{V}\text{ar}\left(\hat{f}_{nh}(x)\right) &= \mathbb{V}\text{ar}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right] \\
&= \frac{1}{n^2h^2}\left\{\sum_{i=1}^{n} \mathbb{V}\text{ar}\left[K\left(\frac{X_i - x}{h}\right)\right] + 0\right\} \\
&= \frac{1}{nh^2}\mathbb{V}\text{ar}\left(K\left(\frac{X_1 - x}{h}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{nh^2}\left\{ \mathbb{E}\left[K^2\left(\frac{X_1 - x}{h}\right)\right] - \left[\mathbb{E}\left(K\left(\frac{X_1 - x}{h}\right)\right)\right]^2 \right\} \\
&= \frac{1}{nh^2}\left\{ \int_{-\infty}^{\infty} f(X_1) K^2\left(\frac{X_1 - x}{h}\right) dX_1 - \left[\int_{-\infty}^{\infty} f(X_1) K\left(\frac{X_1 - x}{h}\right) dX_1\right]^2 \right\} \\
&= \frac{1}{nh^2}\left\{ h\int_{-\infty}^{\infty} f(x + hv) K^2(v) dv - \left[h\int_{-\infty}^{\infty} f(x + hv) K(v) dv\right]^2 \right\} \\
&= \frac{1}{nh^2}\left\{ h\int_{-\infty}^{\infty} [f(x) + f^{(1)}(\xi) hv] K^2(v) dv - O(h^2) \right\} \\
&= \frac{1}{nh}\left\{ f(x)\int_{-\infty}^{\infty} K^2(v) dv + O\left(h\int_{-\infty}^{\infty} |v| K^2(v) dv\right) - O(h) \right\} \\
&= \frac{1}{nh}\{\kappa f(x) + O(h)\},
\end{aligned}
$$

where $\kappa = \int K^2(v) dv$. Then

$$
\begin{aligned}
\mathrm{MSE}(\hat{f}_{nh}(x)) &= \frac{1}{nh}\{\kappa f(x) + O(h)\} + \left[\frac{h^2}{2} f^{(2)}(x)\kappa_2 + O(h^3)\right]^2 \\
&= \frac{h^4}{4}\left[\kappa_2 f^{(2)}(x)\right]^2 + \frac{\kappa f(x)}{nh} + O(h^5 + (n^{-1}h^0)) \\
&= \frac{h^4}{4}\left[\kappa_2 f^{(2)}(x)\right]^2 + \frac{\kappa f(x)}{nh} + o(h^4 + (nh)^{-1}) \\
&= O(h^4 + (nh^{-1})),
\end{aligned}
$$

which concludes the proof. □

In order to prove the convergence in probability $\hat{f}_{nh}(x)$, we will rely on following definitions and theorem.

Definition A.2. *(Order in Probability: Big $O_p(\cdot)$ and Small $o_p(\cdot)$) A sequence of real (possibly vector-valued) random variables $\{X_n\}_{n=1}^{\infty}$ is said to be bounded in probability if, for every $\epsilon > 0$, there exists a constant M and a positive integer N (usually*

$M = M_\epsilon$ and $N = N_\epsilon$), such that

$$\mathbb{P}\left[||X_n|| > M\right] \leq \epsilon, \tag{A.4}$$

for all $n > N$.

That is, we say that X_n is bounded in probability if, for any arbitrary small positive number ϵ, we can always find a positive constant M such that the probability of the absolute value (or norm) of X_n being larger than M is less than ϵ.

Equation (A.4) can be equivalently written as

$$\mathbb{P}\left[||X_n|| \leq M\right] > 1 - \epsilon,$$

for all $n \geq N$ and we write $X_n = O_p(1)$ to indicate that X_n is bounded in probability.

Definition A.3. *(Convergence in Probability) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of real random variables (possibly a finite dimensional vector or matrix-valued), and let X be a random variable having the same dimension as X_n, we say that X_n converge to X in probability if for every (small) $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| < \varepsilon\right) = 1.$$

We use $X_n \xrightarrow{P} X$ to indicate that X_n converges to X in probability and write $X_n = o_p(1)$ if $X_n \xrightarrow{P} 0$.

Theorem A.3. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of real (possibly vector-valued) random variables, and let a_n and b_n be sequences of some non-stochastic, non-negative*

numbers. Then

(*i*) *If* $\mathbb{E}[||X_n||] = O(a_n)$, *then* $X_n = O_p(a_n)$.

(*ii*) *If* $\mathbb{E}[||X_n||^2] = O(b_n)$, *then* $X_n = O_p(b_n^{1/2})$.

Proof. (*i*) From $\mathbb{E}[||X_n||] = O(a_n)$, we know that $\mathbb{E}[||X_n/a_n||] \leq M_0$, for some $M_0 > 0$. For any $\epsilon > 0$, choose $M = M_0/\epsilon$(a finite positive constant). Then by Markov's inequality, we have $\mathbb{P}(||X_n/a_n|| > M) < \frac{\mathbb{E}[||X_n/a_n||]}{M} \leq \epsilon$, which means $||X_n/a_n|| = O_p(1)$ or $||X_n|| = O_p(a_n)$.

(*ii*) From $\mathbb{E}[||X_n||^2] = O(b_n)$, we know that $\mathbb{E}[||X_n^2/b_n||] \leq M_0$, for some $M_0 > 0$. For any $\epsilon > 0$, choose $M = M_0/\epsilon$(a finite positive constant). Then by Markov's inequality, we have $\mathbb{P}(||X_n/b_n^{1/2}|| > M^{1/2}) < \frac{\mathbb{E}[||X_n^2/b_n||]}{M} \leq \epsilon$, which means $||X_n/b_n^{1/2}|| = O_p(1)$ or $||X_n|| = O_p(b_n^{1/2})$. □

By using theorem A.3 (*ii*) and Theorem 2.2, we have

$$\hat{f}_{nh}(x) - f(x) = O_p(h^2 + (nh)^{-1/2}) = o_p(1),$$

thus $\hat{f}_{nh}(x)$ is a consistent estimator for $f(x)$.

Proof. From the definition, one can write

$$\begin{aligned}
\hat{f}_{nh}(x) - f(x) = O_p(h^2 + (nh)^{-1/2}) &= \mathbb{P}\left[\frac{\left|\hat{f}(x) - f(x)\right|}{h^2 + (nh)^{-1/2}} < M\right] > 1 - \varepsilon \\
&= \lim_{nh\to\infty, h\to 0} \mathbb{P}\left(\left|\hat{f}(x) - f(x)\right| < M\left[h^2 + (nh)^{-1/2}\right]\right) > 1 - \varepsilon \\
&= \lim_{nh\to\infty, h\to 0} \mathbb{P}\left(\left|\hat{f}(x) - f(x)\right| < M\left[h^2 + (nh)^{-1/2}\right]\right) = 1
\end{aligned}$$

$$= o_p(1),$$

which concludes the proof. □

A.5 Proof of Bernstein Estimators Properties

In order to prove Theorem 3.3, some intermediate results are, and are given bellow.

Lemma 2. *(Leblanc, 2012b, Lemma 1) Assuming F is continuous (and bounded) and admits two continuous and bounded derivatives on $[0,1]$ and let*

$$T_{j,m}(x) = m^{-j} \sum_{k=0}^{m} (k - mx)^j P_{m,k}(x).$$

Then, the following results are valid. For all $x \in [0,1]$,

$$T_{0,m}(x) = 1, \quad T_{1,m}(x) = 0, \quad T_{2,m}(x) = m^{-1}x(1-x).$$

Proof of Lemma 2. We have

$$\begin{aligned}
T_{0,m}(x) &= m^{-0} \sum_{k=0}^{m} (k - mx)^0 P_{m,k}(x) = 1, \\
T_{1,m}(x) &= \frac{1}{m} \sum_{k=0}^{m} (k - mx) \frac{m!}{k!(m-k)!} x^k (1-x)^{m-k} \\
&= \frac{1}{m} \left[\sum_{k=0}^{m} \left(\frac{m!}{(k-1)!(m-k)!} x^k (1-x)^{m-k} - \frac{m \cdot m!}{k!(m-k)!} x^{k+1} (1-x)^{m-k} \right) \right] \\
&= x \left[\sum_{k=0}^{m} \left(\frac{(m-1)!}{(k-1)!(m-k)!} x^{k-1} (1-x)^{m-k} - \frac{m!}{k!(m-k)!} x^k (1-x)^{m-k} \right) \right] \\
&= 0,
\end{aligned}$$

$$T_{2,m}(x) = m^{-1}x(1-x),$$

which completes the proof. □

Lemma 3. *(Leblanc, 2012a, Lemma 2) Let* $\psi_1(x) = [4\pi x(1-x)]^{-1/2}$ *and* $\psi_2(x) = [x(1-x)/(2\pi)]^{1/2}$. *We define*

$$S_m(x) = \sum_{k=0}^{m} P_{m,k}^2(x),$$

and, for $j = 0, 1$ *and* 2,

$$R_{j,m}(x) = m^{-j} \sum\sum_{0 \le k < l \le m} (k - mx)^j P_{m,k}(x) P_{m,l}(x),$$

then the following results hold:

(i) $0 \le S_m(x) \le 1$ *for* $x \in [0,1]$,

(ii) $S_m(x) = m^{-1/2}[\psi_1(x) + o(1)]$ *for* $x \in (0,1)$,

(iii) $S_m(0) = S_m(1) = 1$,

(iv) $R_{1,m}(x) = m^{-1/2}[-\psi_2(x) + o(1)]$ *for* $x \in (0,1)$,

(v) $0 \le R_{2,m}(x) \le (4m)^{-1}$ *for* $x \in (0,1)$,

(vi) $R_{j,m}(0) = R_{j,m}(1) = 0$ *for* $j = 0, 1, 2$.

Proof of Lemma 3. First note that (i), and (vi) trivially hold. We turn to prove (ii), (iii), (iv) and (v). We follow the proof of Babu et al. (2002, Lemma 3.1) for (ii) based on following theorem.

Theorem A.4. *(Feller (1971)) If F is a lattice distribution with span h, then as $n \to \infty$*

$$\frac{\sqrt{n}}{h} p_n(x) - \phi(x) \to 0$$

uniformly in x. Where $\phi(x)$ is the standard normal density function and $p_n(x) = P\left(\frac{S_n}{\sqrt{n}} = x\right)$ with S_n denotes n times summation of independent random variables $X_1, \ldots, X_n$ identically distributed as F such that

$$\mathbb{E}(X_1) = 0, \qquad \mathbb{V}\mathrm{ar}(X_1) = 1.$$

Note that a lattice distribution is a distribution F such that the random variables X is restricted to values of the form $b, b \pm h, b \pm 2h, \ldots$. Then let U_i, W_j, $i, j = 1, \ldots, m$ be *i.i.d.* Bernoulli random variables with $P(U_1 = 1) = x = 1 - P(U_1 = 0)$, and let $R_i = (U_i - W_i)/\sqrt{2x(1-x)}$. Then we have $\mathbb{E}(R_i) = 0$ and $\mathbb{V}\mathrm{ar}(R_i) = 1$, and

$$S_m(x) = \sum_{k=0}^{m} P_{m,k}^2(x) = \mathbb{P}\left(\sum_{i=1}^{m} U_i = \sum_{i=1}^{m} W_i\right) = \mathbb{P}\left(\sum_{i=1}^{m} R_i = 0\right),$$

because the number of events happened is equal in the two m times Bernoulli experiments. Notice that R_i is a lattice distribution with span $\sqrt{2x(1-x)}$ and we apply Theorem A.4 to have

$$\frac{\sqrt{m}}{\left(\sqrt{2x(1-x)}\right)^{-1}} \mathbb{P}\left(\frac{1}{\sqrt{m}} \sum_{i=1}^{m} R_i = 0\right) \to \phi(0) = \frac{1}{\sqrt{2\pi}},$$

then

$$S_m(x) = P\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{m} R_i = 0\right)$$
$$= m^{-1/2}[\psi_1(x) + o(1)].$$

And for (iii), since $0^0 = 1$,

$$S_m(0) = S_m(1) = 1.$$

The proof for (iv) and (v), one can find them at Leblanc (2012a, Lemma 2). □

Now we turn to proof Theorem 3.3.

Proof of Theorem 3.3.

(i) We begin by calculating the bias

$$\begin{aligned}
\mathbb{E}[\widehat{F}_{n,m}(x)] &= \mathbb{E}\left[\sum_{k=0}^{m} F_n(k/m)P_{m,k}(x)\right] \\
&= \sum_{k=0}^{m} F(k/m)P_{m,k}(x) \\
&= \sum_{k=0}^{m}\left[F(x) + F'(x)(k/m - x) + \frac{F''(x)}{2!}(k/m - x)^2 + o(k/m - x)^2\right]P_{m,k}(x) \\
&= F(x) + F'(x)T_{1,m}(x) + \frac{F''(x)}{2!}T_{2,m}(x) + o(T_{2,m}(x)).
\end{aligned}$$

From the Lemma 2, we can get

$$\mathbb{E}[\widehat{F}_{n,m}(x)] = F(x) + (2m)^{-1}x(1-x)F''(x) + o(m^{-1}),$$

thus,

$$\mathbb{B}\text{ias}[\widehat{F}_{n,m}(x)] = \mathbb{E}[\widehat{F}_{n,m}(x)] - F(x) = m^{-1}b(x) + o(m^{-1}),$$

where $b(x) = 2^{-1}x(1-x)F''(x)$.

(ii) Then we take a look at the variance,

$$\begin{aligned}
\widehat{F}_{n,m}(x) - \mathbb{E}[\widehat{F}_{n,m}(x)] &= \sum_{k=0}^{m}[F_n(k/m) - F(k/m)]P_{m,k}(x) \\
&= \sum_{k=0}^{m}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \le k/m) - F(k/m)\right]P_{m,k}(x) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=0}^{m}(\mathbb{I}(X_i \le k/m) - F(k/m))P_{m,k}(x)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_{i,m},
\end{aligned}$$

where $Y_{i,m} = \sum_{k=0}^{m}(\mathbb{I}(X_i \le k/m) - F(k/m))P_{m,k}(x)$.

Since

$$\mathbb{V}\text{ar}(\widehat{F}_{n,m}(x)) = \mathbb{V}\text{ar}(\widehat{F}_{n,m}(x) - \mathbb{E}[\widehat{F}_{n,m}(x)]) = \frac{1}{n}\mathbb{V}\text{ar}(Y_{1,m}),$$

then we calculate the $\mathbb{V}\text{ar}(Y_{1,m})$, we have

$$\mathbb{E}[Y_{1,m}] = \mathbb{E}\left[\sum_{k=0}^{m}(\mathbb{I}(X_1 \le k/m) - F(k/m))P_{m,k}(x)\right] = 0,$$

and

$$\begin{aligned}
\mathbb{E}[Y_{1,m}^2] &= \mathbb{E}\left[\left(\sum_{k=0}^{m}(I(X_1 \le k/m) - F(k/m))P_{m,k}(x)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{m}(I(X_1 \le k/m) - F(k/m))^2 P_{m,k}^2(x)\right]
\end{aligned}$$

$$+\mathbb{E}\left[2\sum_{0\le k<\ell\le m}\sum(\mathbb{I}(X_1\le k/m)-F(k/m))(\mathbb{I}(X_1\le \ell/m)-F(\ell/m))P_{m,k}(x)P_{m,\ell}(x)\right].$$

Since then,

$$\begin{aligned}\mathbb{E}[(\mathbb{I}(X_1\le k/m)-F(k/m))^2] &= \mathbb{E}[\mathbb{I}^2(X_1\le k/m)+F(k/m)^2-2F(k/m)\mathbb{I}(X_1\le k/m)]\\ &= \mathbb{E}[\mathbb{I}(X_1\le k/m)]+F(k/m)^2-2F(k/m)\mathbb{E}[\mathbb{I}(X_1\le k/m)]\\ &= F(k/m)-F(k/m)^2,\end{aligned}$$

thus,

$$\mathbb{E}\left[\sum_{k=0}^{m}(\mathbb{I}(X_1\le k/m)-F(k/m))^2P_{m,k}^2(x)\right]=\sum_{k=0}^{m}F(k/m)P_{m,k}^2(x)-\sum_{k=0}^{m}F(k/m)^2P_{m,k}^2(x).$$

Besides, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{0\le k<\ell\le m}\sum(\mathbb{I}(X_1\le k/m)-F(k/m))(\mathbb{I}(X_1\le \ell/m)-F(\ell/m))\right]\\ &\quad=\mathbb{E}[\mathbb{I}(X_1\le k/m)I(X_1\le \ell/m)]+\mathbb{E}[F(k/m)F(\ell/m)]-\mathbb{E}[F(\ell/m)\mathbb{I}(X_1\le k/m)]\\ &\qquad-\mathbb{E}[\mathbb{I}(X_1\le \ell/m)F(k/m)].\end{aligned}$$

Since $k<\ell$ and $\mathbb{E}[\mathbb{I}(X_1\le k/m)\mathbb{I}(X_1\le \ell/m)]=F(k/m)$, thus

$$\mathbb{E}\left[\sum_{0\le k<\ell\le m}\sum(\mathbb{I}(X_1\le k/m)-F(k/m))(\mathbb{I}(X_1\le \ell/m)-F(\ell/m))\right]=F(k/m)-F(k/m)F(\ell/m),$$

therefore, one can rewrite the $\mathbb{E}[Y_{1,m}^2]$ as

$$\begin{aligned}\mathbb{E}[Y_{1,m}^2] &= \sum_{k=0}^{m} F(k/m)P_{m,k}^2(x) + 2\sum_{0\le k<\ell\le m}\sum [F(k/m) - F(k/m)F(\ell/m)]P_{m,k}(x)P_{m,\ell}(x) \\ &\quad - \sum_{k=0}^{m} F(k/m)^2 P_{m,k}^2(x) \\ &= \sum_{k=0}^{m} F(k/m)P_{m,k}^2(x) + 2\sum_{0\le k<\ell\le m}\sum F(k/m)P_{m,k}(x)P_{m,\ell}(x) \\ &\quad - \left[\sum_{k=0}^{m} F(k/m)P_{m,k}(x)\right]^2. \end{aligned} \tag{A.5}$$

By Taylor expansion, $F(k/m) = F(x) + O(|k/m - x|)$, then we have

$$\sum_{k=0}^{m} F(k/m)P_{m,k}^2(x) = F(x)S_m(x) + O(I_m(x)),$$

where $I_m(x) = \sum_{k=0}^{m} |k/m - x|P_{m,k}^2(x)$. For the second term of (A.5), we can rewrite F(k/m) as

$$F(k/m) = F(x) + (k/m - x)F'(x) + O((k/m - x)^2),$$

and note that

$$1 = \sum_{k=0}^{m}\sum_{l=0}^{m} P_{m,k}(x)P_{m,\ell}(x) = 2R_{0,m}(x) + S_m(x),$$

then

$$R_{0,m}(x) = \frac{1}{2}[1 - S_m(x)].$$

Thus, according to Lemma 3 (v)

$$\sum_{0\le k<l\le m}\sum F(k/m)P_{m,k}(x)P_{m,\ell}(x) = F(x)R_{0,m}(x) + F'(x)R_{1,m}(x) + O(R_{2,m}(x))$$
$$= \frac{1}{2}F(x)[1 - S_m(x)] + F'(x)R_{1,m}(x) + O(m^{-1}).$$

By then, we denote $\sum_{k=0}^{m} F(k/m)P_{m,k} = B_m(x)$, one can write

$$\mathbb{E}[Y_{1,m}^2] = F(x) + 2F'(x)R_{1,m}(x) + O(m^{-1}) + O(I_m(x)) - B_m^2(x).$$

By using Lemma 3 (iv), we have

$$\mathbb{E}[Y_{1,m}^2] = F(x) - B_m^2(x) - m^{-1/2}V(x) + O(m^{-1}) + O(I_m(x)),$$

where $V(x) = F'(x)[2x(1-x)/\pi]^{1/2}$.

By using Cauchy-schwarz inequality and Lemma 3 (ii), we have

$$I_m(x) = \sum_{k=0}^{m} |k/m - x|P_{m,k}^2(x) \le \left[\sum_{k=0}^{m} P_{m,k}^3(x)\sum_{k=0}^{m} P_{m,k}(x)(k/m - x)^2\right]^{1/2}$$
$$= \left(T_{2,m}(x)\sum_{k=0}^{m} P_{m,k}^3(x)\right)^{1/2}$$
$$= O\left(\frac{1}{m}\sum_{k=0}^{m} P_{m,k}^3(x)\right)^{1/2}$$
$$= O(m^{-3/4}),$$

where $\sum_{k=0}^{m} P_{m,k}^3(x) = O(m^{-1/2})$ by the same operation in Lemma 3 (ii).

Then,

$$\begin{aligned}\mathbb{E}[Y_{1,m}^2] &= F(x) - B_m^2(x) - m^{-1/2}V(x) + o(m^{-1/2}) \\ &= \sigma^2(x) - m^{-1/2}V(x) + o(m^{-1/2})\end{aligned}$$

where $V(x) = f(x)[2x(1-x)/\pi]^{1/2}$ and $\sigma^2(x) = F(x)[1-F(x)]$. Thus,

$$\mathbb{V}\text{ar}[\widehat{F}_{n,m}(x)] = n^{-1}\sigma^2(x) - n^{-1}m^{-1/2}V(x) + o(n^{-1}m^{-1/2}).$$

(iii) Finally, we get that

$$\text{MSE}[\widehat{F}_{n,m}(x)] = n^{-1}\sigma^2(x) - n^{-1}m^{-1/2}V(x) + m^{-2}b^2(x) + o(m^{-2}) + o(n^{-1}m^{-1/2}),$$

which concludes the proof. □

Bibliography

Azzalini, A. and A. W. Bowman (1990). A look at some data on the old faithful geyser. *Applied Statistics*, 357–365.

Babu, G. J., A. J. Canty, and Y. P. Chaubey (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planing and Inference. 105*, 377–392.

Babu, G. J. and Y. P. Chaubey (2006). Smooth estimation of a distribution and density function on a hyper-cube using Bernstein polynomials for dependent random vectors. *Statistics and Probability Letters 76*, 959–969.

Bashtannyk, D. M. and R. J. Hyndman (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics and Data Analysis 36*, 279–298.

Belalia, M. (2016). On the asymptotic properties of the bernstein estimator of the multivariate distribution function. *Statistics & Probability Letters 110*(C), 249–256.

Belalia, M., T. Bouezmarni, and A. Leblanc (2017). Smooth conditional distribution estimators using bernstein polynomials. *Computational Statistics and Data Analysis 111*, 166 – 182.

Belalia, M., T. Bouezmarni, and A. Leblanc (2019). Bernstein conditional density estimation with application to conditional distribution and regression functions. *Journal of the Korean Statistical Society 48*(3), 356–383.

Di Lucca, M. A., A. Guglielmi, P. Müller, and F. A. Quintana (2013). A simple class of bayesian nonparametric autoregression models. *Bayesian Anal. 8*(1), 63–88.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *Annals of Statistics 35*(6), 2504–2535.

Efromovich, S. (2010). Oracle inequality for conditional density estimation and an actuarial example. *Annals of the Institute of Statistical Mathematics 62*(2), 249–275.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Fan, J. and T. H. Yim (2004). A crossvalidation method for estimating conditional densities. *Biometrika 91*(4), 819–834.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. New York: John Wiley & Sons Inc.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Annals of Statistics 28*, 1264–1280.

Hall, P., J. Racine, and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association 99*(468), 1015–1026.

Hall, P., R. C. L. Wolff, and Q. Yao (1999). Methods for estimating a conditional distribution function. *Journal of the american Statistical Association 94*(445), 154–163.

Hansen, B. E. (2004). Nonparametric estimation of smooth conditional distributions. Technical report, University of Wisconsin, Madison.

Hyndman, R. J., D. M. Bashtannyk, and G. K. Grunwald (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics 5*(4), 315–336.

Hyndman, R. J. and Q. Yao (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Journal of Nonparametric Statistics 14*(3), 259–278.

Leblanc, A. (2009). Chung-Smirnov property for Bernstein estimators of distribution functions. *Journal of Nonparametric Statistics 22*(2), 459–475.

Leblanc, A. (2010). A bias-reduced approach to density estimation using Bernstein polynomials. *Journal of Nonparametric Statistics 22*, 459–475.

Leblanc, A. (2012a). On estimating distribution functions using Bernstein polynomials. *Annals of the Institute of Statistical Mathematics 64*, 919–943.

Leblanc, A. (2012b). On the boundary properties of Bernstein polynomial estimators of density and distribution functions. *Journal of Statistical Planing and Inference 142*, 2762–2778.

Li, Q. and J. S. Racine (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Lorentz, G. (1986). *Bernstein Polynomials* (2nd ed.). New York: Chelsea Publishing.

Matzner-Løber, E., A. Gannoun, and J. G. De Gooijer (1998). Nonparametric forecasting: A comparison of three kernel-based methods. *Commun. Stat. - Theory Methods 27*(7), 1593–1617.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications 9*(1), 141–142.

Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theory of Probability and its Applications 10*, 186–190.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics 27*, 105–126.

Petrone, S. (1999b). Random bernstein polynomials. *Scandinavian Journal of Statistics 26*, 373–393.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics 27*(3), 832–837.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In P. Krishnaiah (Ed.), *Multivariate Analysis II*, New York, pp. 25–31. Academic Press.

Serfling, R. J. (2002). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Routledge.

Stone, C. J. (1977). Consistent nonparametric regression. *The annals of statistics*, 595–620.

Takeuchi, I., K. Nomura, and T. Kanamori (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation 21*(2), 533–559.

Tenbusch, A. (1994). Two-dimensional Bernstein polynomial density estimation. *Metrika 41*, 233–253.

Veraverbeke, N., I. Gijbels, and M. Omelka (2014). Preadjusted non-parametric estimation of a conditional distribution function. *Journal of the Royal Statistical Society: Series B 76*, 399–438.

Vitale, R. (1975). A Bernstein polynomial approach to density estimation. In M. L. Puri (Ed.), *Statistical Inference and Related Topics*, Volume 2, New York, pp. 87–99. Academic Press.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya A 26*(4), 359–372.

Xian, S. (2005). Kernel smoothing based on bernstein polynomials. Master's thesis, University of Manitoba.

Vita Auctoris

Mr. Guanjie Lyu was born in 1996, Dazhou, Sichuan, China. He graduated from Shijiazhuang Railway University in 2018 with a Bachelor of Science degree. For a peruse of continuous education, he came to Canada. He is currently a candidate for the Master of Science degree in Statistics at the University of Windsor.